System Acceptance Report for NSF award 1445604 "High Performance Computing System Acquisition: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment"

Craig A. Stewart[1], PI
David Y. Hancock[1], Systems Lead
Matthew Vaughn[2], Co-PI
Nirav Merchant[3], Co-PI
J. Michael Lowe[1], Jetstream lead sysadmin
Jeremy Fischer[1], Senior Technical Advisor
Lee Liming[4], Globus Services Lead and SI
James Taylor[5], Jetstream Co-PI and Galaxy PI
Enis Afgan[5], Galaxy implementation lead
George Turner[1], System Architect
C. Bret Hammond[1], Jetstream sysadmin
Edwin Skidmore[3], Atmosphere software lead
Michael Packard[2], Senior systems administrator
Ian Foster[4], Co-PI

[1]Indiana University Pervasive Technology Institute
[2]University of Texas at Austin Texas Advanced Computing Center
[3]University of Arizona
[4]University of Chicago Computation Institute
[5]Johns Hopkins University

May 11, 2016

Stewart, C.A., Hancock, D.Y., Vaughn, M., Merchant, N., Lowe, J.M., Fischer, J., Liming, L., Taylor, J., Afghan, E., Hammond, C.B., Skidmore, E., Foster, I. "System Acceptance Report for NSF award 1445604 "High Performance Computing System Acquisition: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment"" Indiana University, Bloomington, IN. PTI Technical Report PTI-TR16-003 May 11, 2016.

Table of Contents

This material is based upon work supported in part by the National Science Foundation under Award 1445604 ”High Performance Computing System Acquisition: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment.” Partners in the Jetstream Implementation include the Indiana University Pervasive Technology Institute; University of Texas at Austin Texas Advanced Computing Center; University of Arizona; University of Chicago Computation Institute; Johns Hopkins University.

The Indiana University Pervasive Technology Institute has also supported Jetstream implementation and related activities. The IU Pervasive Technology Institute is supported by Indiana University and has received major support from the Lilly Endowment, Inc.

1. Executive summary

The Jetstream project began as a direct result of NSF solicitation 14-536, which states in part:

> *The intent of this solicitation is to request proposals from organizations willing to serve as Resource Providers within the NSF eXtreme Digital (XD) program. The current solicitation is intended to complement previous NSF investments in advanced computational infrastructure by exploring new and creative approaches to delivering innovative computational resources to an increasingly diverse community and portfolio of scientific research and education.*

In our proposal to the NSF, which subsequently resulted in NSF award 1445604, we proposed the Jetstream system which provides:

- "Self-serve" academic cloud services, enabling researchers or students to select a VM image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing. Authentication to this "self-serve" environment is via Globus using XSEDE credentials.
- Hosting for persistent Science Gateways. Jetstream supports persistent science gateways, including the capability of hosting persistent science gateways within a VM when the nature of the gateway is consistent with operation within a VM.
    - Galaxy is one of the initial science gateways supported.
- Data movement, storage and dissemination.
    - Jetstream supports data transfer with Globus Connect.
    - Users are able to store VMs in the Indiana University persistent digital repository, IUScholarWorks (scholarworks.iu.edu) and obtain a Digital Object Identifier (DOI) that is associated with the VM stored.
- Virtual Linux desktop services delivered from Jetstream to tablet devices. This service is aimed at increasing access to Jetstream for users at institutions with limited resources including small schools, schools in EPSCoR states, and Minority Serving Institutions.

In this document, we demonstrate that:

- The hardware components purchased and installed as Jetstream match the hardware component numbers and capabilities as specified in the system description in the final grant proposal revised scope of work (based on the final budget specified by the NSF).
- Jetstream meets the requirements set out in NSF solicitation 14-536 in terms of integration with XSEDE.
- Jetstream fulfills the tests specified in our peer-reviewed Project Execution Plan in ways that demonstrate that Jetstream as currently implemented is indeed the system proposed, and it operates successfully.
- Jetstream is being allocated as called for in NSF solicitation 14-536 as a resource at 90% of its capacity via NSF-specified allocation processes operated by XSEDE (the Extreme Science and Engineering Discovery Environment).

The above items fulfill the acceptance tests specified in the Project Execution Plan. However, Jetstream is unusual within the history of NSF acquisitions in several ways. It is the first system intended to serve as a production cloud system for end-user scientists. The NSF has never funded such a system before. As such, Jetstream is a first-of-a-kind acquisition for the NSF, and in this particular sense it is also a pilot project – and a learning experience for the intended user community, the implementing team, and the NSF.

As a resource intended as a production resource, it is worthwhile to also consider our experiences from friendly user and early operations modes. From our friendly user and early operations mode we have learned the following:

- The system is usable – it has a variety of software tools available that make it useful to the user communities identified as intended users of Jetstream.
- Jetstream has been used and people like using it. Jetstream has been used by a total of 287 people, of them 163 "end-user researchers or students" and 124 staff of the Jetstream implementation team and XSEDE (the eXtreme Science and Engineering Environment).
- Jetstream has been used to perform meaningful scientific research. In this document, we have included several short summaries of useful incremental scientific results that have been generated using Jetstream. These include scientific results in several areas of research we targeted as priorities in our initial proposal. Early results involve genomics and field biology, psychology, computer and computational science. Storage facilities offered as part of the integrated suite of Jetstream services are being used to enhance replicability of scientific analyses. Perhaps the strongest sign of the utility of Jetstream is that analyses performed with Jetstream by people not affiliated with the project have already produced useful incremental results that will accelerate the submission of scientific technical reports to peer-reviewed journals. During early operations, Jetstream has also been used in educational activities in courses that receive university course credits and as part of doctoral dissertation research.

Data included here are as of midnight April 30, 2016.

2. Introduction: NSF goals to increase diversity of users of its cyberinfrastructure, purpose of Jetstream, and Jetstream description

The National Science Foundation (NSF) has for decades provided for the national research community computer, storage, visualization, network, and human resources that we now refer to collectively as cyberinfrastructure (CI). These resources have been provisioned through grant solicitations and grant awards to organizations that deliver NSF-funded resources to the research community on behalf of the NSF. One of the most recent major solicitations was NSF - Program Solicitation 14-536 [1], which begins as follows:

> *The intent of this solicitation is to request proposals from organizations willing to serve as Resource [Service] Providers within the NSF eXtreme Digital (XD)*[1] *program. The current solicitation is intended to complement previous NSF investments in advanced computational infrastructure by exploring new and creative approaches to delivering innovative computational resources to an increasingly diverse community and portfolio of scientific research and education.*

The eXtreme Digital (XD) program includes the eXtreme Science and Engineering Discovery Environment (XSEDE), which serves a coordinating and supporting function, and the several NSF-funded Service Providers that provide advanced CI systems for use via XSEDE. In 2013 and 2014, the NSF's solicitations for Service Providers focused on increasing the diversity of CI resources provided to the national research community via XSEDE. This was a result of many factors, ranging from workshops, surveys, and analysis of usage of NSF-funded CI. For example, the NSF estimates that 350,000 researchers, educators, and learners received direct support during the year ending September 2015 [2]. Yet, under 2% of these individuals completed a computation, data analysis, or visualization task on XD program resources and less than 4% had an account on the XSEDE portal [3], [4].

NSF solicitation 14-536 specifically states increasing diversity of users of advanced CI as one of the goals of the solicitation:

> *Consistent with the Advanced Computing Infrastructure: Vision and Strategic Plan (February 2012), the current solicitation is focused on expanding the use of high-end resources to a much larger and more diverse community. To quote from that strategic plan, the goal is to "... position and support the entire spectrum of NSF-funded communities ... and to promote a more comprehensive and balanced portfolio .... to support multidisciplinary computational and data-enabled science and engineering that in turn supports the entire scientific, engineering and educational community." Thus, while continuing to provide essential and needed resources to the more traditional users of HPC, this solicitation expands the horizon to include research communities that are not users of traditional HPC*

[1] We recognize that there is not a formally chartered entity within NSF funded activities called "the XD Program." We will follow the NSF example and use the name "the XD Program" to refer to XSEDE and the NSF-funded resource providers that manage and deliver resources that are allocated and supported via XSEDE under NSF direction so to do.

systems, but who would benefit from advanced computational capabilities at the national level.

The primary purpose of Jetstream is to provide researchers with interactive access to a handful of CPUs, now, whenever "now" is [5]. This is a particular mode of use that has never before been a focus within the XD program, and this sort of use is today best supported in a cloud computing environment.

Jetstream is a first-of-a-kind acquisition for the NSF – the first system intended to be a production cloud for end-user scientists. (The NSF has previously funded three cloud systems all of which were specifically for computer science and computational science research). Jetstream is thus unusual because on the one hand it is intended to function from the user standpoint as a production resource. On the other hand, Jetstream is a first-of-a-kind acquisition for the NSF, and in this particular sense it is also a pilot project – and a learning experience for the intended user community, the implementing team, and the NSF.

Jetstream employs cloud-deployed virtual machine (VM) technology to support, in particular, researchers working in the long tail of science [6] and, in general, add to NSF efforts to expand the range and number of scientists using XD resources. As a cloud resource that enables end users to provision VMs of the users choosing, Jetstream is self-provisioned from the user's standpoint.

Jetstream addresses a clear gap that existed in the collected Service Provider resources available via XSEDE at the time we proposed this system. The current ecosystem includes systems providing scalable High-Performance Computing, large memory, large data, and high-throughput resources and now with Jetstream: interactive cloud-based resources. Jetstream in particular compliments other recent NSF acquisitions intended to increase the diversity of resources in and users of the XD program: Comet, Wrangler, and Bridges.

Building on cloud concepts and software, Jetstream is designed to deliver the services and programming models needed by researchers working in the "long tail of science" and deliver them in a way that is (and is perceived to be) easily accessible and valuable to them. In particular, Jetstream:

- Offers "self-serve" academic cloud services, enabling researchers or students to select a pre-existing VM image or to create a new virtual environment for personalized research computing.
- Hosts persistent Science Gateways.
- Enables data movement, storage, and dissemination.
- Provides virtual desktop services delivered to tablet devices which increases access to CI for users at resource-limited institutions (e.g. small schools, schools in EPSCoR states, and Minority Serving Institutions).

A brief video describing the basic capabilities of Jetstream is online at https://www.youtube.com/watch?v=olo5OFeVzHk.

2.1. Purpose of this document

The purpose of this acceptance report is to present data that demonstrate the following:

- The hardware components purchased and installed as Jetstream match the hardware component numbers and capabilities as specified in the system description in the final grant proposal revised scope of work (based on the final budget specified by the NSF).
- The Jetstream system meets the basic criterion for the system as specified in the original solicitation NSF 14-536 the Cooperative Agreement with the NSF in terms of integration with XSEDE.
- The Jetstream system, as it exists and operates today, is the system described in the Cooperative Agreement between the National Science Foundation and Indiana University for NSF Award 1445604 as defined in the executed cooperative agreement and Project Execution Plan (PEP).
- The Jetstream system fulfills the NSF-mandated criterion of having 90% of its capacity allocable and allocated through the XSEDE-managed allocation process.

In addition to the above steps, which we believe fulfill specific performance tests indicated in the PEP, we present information gleaned from our experiences in early operations of Jetstream as a pilot implementation for the NSF. These early operations experience demonstrate:

- The Jetstream system as currently implemented provides resources of utility to the US science and engineering open research community.
- The Jetstream system has been used to perform analyses of practical use in scientific discovery in ways that will contribute to society's knowledge through additions to the corpus of peer-reviewed and openly published technical reports and papers.

The above demonstrate practical and realized capabilities of Jetstream as a computational resource to support useful increments in the national research community's ongoing activities. These incremental contributions in the brief early operations phase should provide confidence that Jetstream, in an anticipated four-year management and operations phase, will be of significant practical value to the national open science research community.

As an additional step in this experience shared between the NSF, the Jetstream team, and the national user community, we present operational metrics for the month of April 2016 and compare them to the current draft metrics for the Management and Operations phase of the planned Jetstream award. These data demonstrate:

- During the early operations phase Jetstream's delivered capabilities and usage by the national research community meet almost all of the metrics currently set for the post-acceptance Management and Operations phase of Jetstream as a resource for the national community.

In sum, the purpose of this report is to demonstrate that Jetstream as now implemented is indeed the system described in our proposal (as modified in scope of work change statements based on the final budget specified by the NSF), that it fulfills the requirements for acceptance as specified

in the Project Execution Plan, and that the results of early operations activities with Jetstream show that it has potential to aid the US research community and that potential has been realized in the form of useful increments of scientific research that will contribute to acceleration in the publication of reports in peer-reviewed scientific journals.

2.2. Responding to NSF-14-536 – initial partnership and Use cases

The three institutions that initially agreed to collaborate on the response to NSF 14-536 that led to award 1445604 were Indiana University, the University of Texas at Austin, and the University of Chicago. Indiana University (IU) has an established local history of providing advanced computing, storage, visualization, and human resources supporting researchers, scholars, and artists in diverse fields ranging from theatre lighting experts to ethnographers, biologists to physicists, and composers to fine artists. Solicitation 14-536 seemed a natural fit for IU. The Indiana University Pervasive Technology Institute[2] (IUPTI) is a relative newcomer to delivery of national cyberinfrastructure (CI) resources, having received its first funding to deliver computational and storage resources to the national research community in 2003 [7]. Since then, IUPTI has increased the scale and importance of its role in delivering federally funded resources to the national science and engineering research community. IUPTI and Texas Advanced Computing Center (TACC) at the University of Texas at Austin are long-term collaborators, and had at the time of the release of NSF 14-536 already partnered – with TACC in the lead – on a successful proposal to implement the Wrangler system. Involvement of the University of Chicago Computation Institute to the collaboration to provide Jetstream was an extension of the Wrangler partnership, with an expectation that the Computation Institute's involvement would at a minimum involve implementation of Globus-based tools for authentication with XSEDE account management systems (as required by NSF 14-536) and file movement services.

In order to identify and organize our understanding of community needs we employed a formal use case analysis structure [8]–[10]. We developed dozens of domain-specific use case descriptions and several use case descriptions based on mode of use.

The many domain specific use case descriptions we identified and documented may be distilled into one canonical use case with one general proposed solution, as follows:

Generalization of domain-based use cases. "A researcher wants straightforward access to specific, usually interactive tools to analyze data, delivered in a manner congruent with their normal operations and often driven by availability of new data. A tool producer develops new analysis routines and methods to address research bottlenecks and needs to make said tool available to experimentalists without having them contend with technical complexities of operating system and software dependencies." *Proposed solution:* Develop an accessible platform where application creators can easily publish and share within a VM image, and end users can easily invoke runnable instances of these applications via virtualization.

[2] The IUPTI is a collaborative organization within IU led by an Executive Director in the Office of the Vice President for Information Technology (OVPIT) and including subunits of OVPIT, the School of Informatics and Computing, and the Maurer School of Law.

Among the use cases distinguished by mode of access or mode of use of CI resources the following four involve the largest number of users:

Enable analysis of public data sets at small schools with limited CI budgets (including MSIs). There are many highly qualified faculty researchers at small US colleges and universities, including at MSIs, who could do important, original research analyzing data from publicly available data sets that have yet to be fully mined for information and insights. Students could also participate in such research. However, in many cases the ability to obtain and work with this data is limited by local network capability, lack of local computing equipment, and lack of system administrator resources to support research computing. *Proposed solution:* Provide a Jetstream VM image featuring a user-friendly virtual Linux desktop. The virtual desktop will run on Jetstream with screen images delivered to tablet devices on cellular connections or to older PCs on slow networks. The virtual desktop will guide users to information, training materials, and XSEDE and XD program resources, and let users execute data analyses on Jetstream.

Enable use of proprietary software. The critical path for many research analyses includes licensed applications that will make use of modest levels of parallelism. Examples include Mathematica, IDL, and MATLAB. These software packages are not available on most XD resources because of the administrative complexity and expense of providing licenses in a shared environment. *Proposed solution:* Enable Jetstream users to run software such as MATLAB using their own licenses. Based on user demand and vendor cooperation, expand options for using other commercial software on Jetstream**.**

Facilitate reproducible data analyses. There is considerable desire within the scientific community to enable reproducibility of data analyses and published research [11]–[13]. Services such as RunMyCode [14] allow one to disseminate software, data, and scripts, but do not currently provide environments in which to actually run the code. Commercial cloud services can be used, but at indeterminate cost with possibly substantial work on the part of the users. *Proposed solutions:* Enable researchers to easily publish a VM containing their analysis tools, including, in the case of published research, the input data, scripts, and output data in a VM image. Make such VMs downloadable, publishable via services like RunMyCode, or available via a persistent digital archive such as IUScholarWorks [15]. Make VMs easily discoverable by associating a DOI and advertising their metadata via the Globus Data Publication service [16].

Enhance ease of Science Gateway deployment. Science Gateways provide a web-accessible implementation of particular analyses and scientific workflows. While straightforward to use, they can be labor intensive to create. Lacking a generally available, easy-to-follow cookbook, extensive server-side programming is often required to make a science gateway work, and they often involve use of a distributed workflow engine such as Pegasus [17], Taverna [18], Unicore [19], Kepler [20], and Apache Airavata [21]. System administrator intervention is often required to enable use of such software on a local cluster along with changes to network security policy. There are many more groups who maintain or develop XSEDE science gateways, with many working to implement such tools for XSEDE. *Proposed solution:* Provide a gateway builder's toolkit, including VMs with commonly used workflow engines installed and ready to configure and XSEDE tools, coupled with a platform for persistently hosting web services.

2.3. Architectural and support implications of use cases

The use cases we identified clearly called for a solution that can generally be described as "provide a handful of CPU cores to an end user now, whenever now is, interactively." Even in terms of supporting Science Gateway deployment, the current challenges seem greater in provisioning the interactive front of gateways than the sometimes massive supercomputers that constitute behind-the-scenes resources. A cloud-based solution was obvious. Given that, and budget guidance in the solicitation, basing a cloud on the OpenStack cloud software environment was similarly obvious. We fairly quickly settled on a delivery strategy based on interactive activation and delivery of VMs – similar to Amazon Web Services (AWS) or Microsoft Azure, but customized for science. Discussions with potential users suggested that neither the AWS, Azure, nor native OpenStack interfaces would be viewed as sufficiently user friendly by the researchers we intended to support. We therefore selected, as the central feature of the user experience, the already popular and successful Atmosphere cloud interface and orchestration layer. Atmosphere, developed by the University of Arizona, is an intuitive user interface combined with powerful cloud service management and orchestration capabilities. Therefore, we added the University of Arizona to the collaboration and defined our basic architectural response to the community needs we would propose to the NSF under solicitation 14-536: a system that delivers VMs, interactively, to end users providing a modest number of processor cores and modest computational power, delivered in an environment that supported a variety of modes of access and enhanced reproducibility of analyses. Our belief was that we could implement a system that over the course of its life would provide resources for several thousand users.

2.4. Support strategy and community involvement

Resource Providers funded by the NSF to deliver services within the NSF XD program work cooperatively with the NSF-funded XSEDE project (eXtreme Science and Engineering Discovery Environment). Service Providers are funded in typically two phases: initial funding to implement a resource, and management and operations of the resource after the NSF has formally accepted it. The management & operations (M&O) formula used by the NSF is that annual management and operations funding for a system is set at 20% of the cost of acquiring the system. The M&O formula is not sufficient to support, manage, and operate a system independent of the services provided by XSEDE. As partners in XSEDE, IUPTI, TACC, and UC all understand that, as of the time of the release of NSF 14-536, there was insufficient funding within the XSEDE budget to provide support to several thousand individual researchers and students. (We now also know that there will be budget cuts relative to XSEDE in the currently proposed successor XSEDE2.)

The core leadership team of what would later be called Jetstream decided early on that the only practicable way to significantly increase the diversity and number of users of XD Program resources, with a solicitation having a maximum of $12M total in funding specified, was to apply the concepts of the leveraged support model developed by IU in the 1990s [22], [23]. The leveraged support model can be described briefly as "support the supporters, leverage online support tools, and reserve the use of expert human consultants for very challenging problems without pre-existing documented solutions." As obvious as this sounds today, this was ground breaking when IU began this in the 1990s. As applied to Jetstream, this approach suggested that

we engage leaders and aggregators of communities of scientists and engineers working in areas with needs not well met by the XD Program. We thus worked extensively with leaders of virtual organizations (VOs), Communities of Practice (CoPs), disciplinary groups (e.g. "quantitative social scientists" or "field biologists"), or other groupings that have some important factor in common. Examples of the latter include colleges and universities with high quality faculty but strong funding limitations, e.g. some HBCUs, some institutions in EPSCoR states, tribal colleges, or researchers who would like to use MATLAB on an XD Program resource. Our support model for Jetstream thus became to depend on XSEDE, as much as XSEDE is able to help in support of Jetstream, but focus on existing VO and community support and information exchange structures in order to support large numbers of scientists and engineers with CI needs different than the needs that XSEDE was well experienced in meeting at large scale.

This support strategy led to the addition of Johns Hopkins University and an additional role for the University of Arizona. These two universities lead and represent large user communities (Galaxy and iPlant, respectively). In both cases these communities are notable for the disparity between the large number of users and significant investment by NSF in software development and implementation as contrasted with modest NSF support for hardware infrastructure. More succinctly: as of 2014 there were no two user communities with more researchers using advanced CI tools but not using XD resources than the communities of iPlant and Galaxy users.

2.5. Purpose

Based on hundreds of hours of interviews with potential users, discussions with collaborators, and constant return to the goals set out in NSF solicitation 14-536 [1], we arrived at the following statement of Jetstream as a computational resource:

The purpose of the Jetstream computational resource is to ensure that the science and engineering community has ready access to the advanced computational and data-driven capabilities required to tackle today's most complex problems and issues. Jetstream will in particular complement previous NSF investments in advanced computational infrastructure by adding its first cloud environment for use in science and engineering research across all areas of research and education supported by the NSF. Jetstream will be a new type of computational research resource for the national open (unclassified) research community - a data analysis and computational resource that US scientists and engineers will use interactively. This system will enable many US researchers and engineers to make new discoveries that are important to understanding the world around us and will help researchers make new discoveries that improve the quality of life of American citizens.

2.6. Project vision

The vision for the Jetstream project is that Jetstream will be a managed science and engineering cloud – a cloud managed and operated in order to support open science and engineering research in the US.

2.7. Project mission

The mission of the Jetstream project is to provide an interactive, on-demand cloud-based computational system that allows researchers to analyze their data "now" – whenever now is – aimed particularly at researchers working in the "long tail of science."

Jetstream as a facility and the Jetstream team as a management and support group will complement existing NSF-funded cyberinfrastructure resources supported by XSEDE (the eXtreme Science and Engineering Discovery Environment). In particular, Jetstream aims to provide resources that may be used interactively at any time of day or night when a handful of processor cores are needed, and provide large-scale computational use during "non-peak" hours via the API for the Atmosphere user interface. The Jetstream team's objective is that the system be known first and foremost for the distinctive research results and training outcomes it has enabled.

2.8. Description of the project deliverables

The project deliverable for the construction phase of NSF award 1445604 is the Jetstream system. Jetstream is a configurable large-scale computing resource that leverages both on-demand and persistent virtual machine technology to support a much wider array of software environments and services than current NSF resources can accommodate. As a fully configurable "cloud" resource, Jetstream bridges the obvious major gap in the current ecosystem, which has machines targeted at large-scale High-Performance Computing, high memory, large data, high-throughput, and visualization resources. As the open cloud for science, Jetstream provides:

- "Self-serve" academic cloud services, enabling researchers or students to select a VM image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing. Authentication to this "self-serve" environment is via Globus using XSEDE credentials.
- Hosting for persistent Science Gateways. Jetstream supports persistent science gateways, including the capability of hosting persistent science gateways within a VM when the nature of the gateway is consistent with operation within a VM. Galaxy is one of the initial science gateways supported.
- Data movement, storage and dissemination.
    - Jetstream supports data transfer with Globus Connect.
    - Users are able to store VMs in the Indiana University persistent digital repository, IUScholarWorks (scholarworks.iu.edu) and obtain a Digital Object Identifier (DOI) that is associated with the VM stored.
- Virtual Linux desktop services delivered from Jetstream to tablet devices. This service is aimed at increasing access to Jetstream for users at institutions with limited resources including small schools, schools in EPSCoR states, and Minority Serving Institutions.

Two papers regarding the capabilities of and plans for Jetstream have already been published (see [24][25]).

2.9. Project Execution Plan and acceptance criteria

The Project Execution Plan for Jetstream was submitted to the NSF, sent out for peer review, and then revised by the project team. It has been updated as time and experiences warrant. The PEP specifies acceptance criteria. Throughout this document we will refer to the PEP and indicate whether tests have been passed or not. All acceptance test results are presented in this document using easy-to-interpret visual symbols to indicate status. These symbols are shown below in Table 1 (they are adapted with gratitude from XSEDE reports [26]):

Table 1. Status icons representing outcomes of performance tests

Definition	Icon
Outcome is successful and complete	G
Outcome is in progress but not yet fulfilled or achieved	Y
Outcome is unsuccessful	R
Outcome is incomplete or metrics aren't available	W

2.10. System description

Jetstream is a physically distributed cloud with three hardware components: 1) a test system is located at the University of Arizona, 2) one production system at the Indiana University Pervasive Technology Institute, and 3) one production system (identical to the system at IU) at the Texas Advanced Computing Center at the University of Texas at Austin. The two production systems are tied to the XSEDE network at 10 Gbps and to the Internet2 backbone at 100 Gbps (see Figure 1).

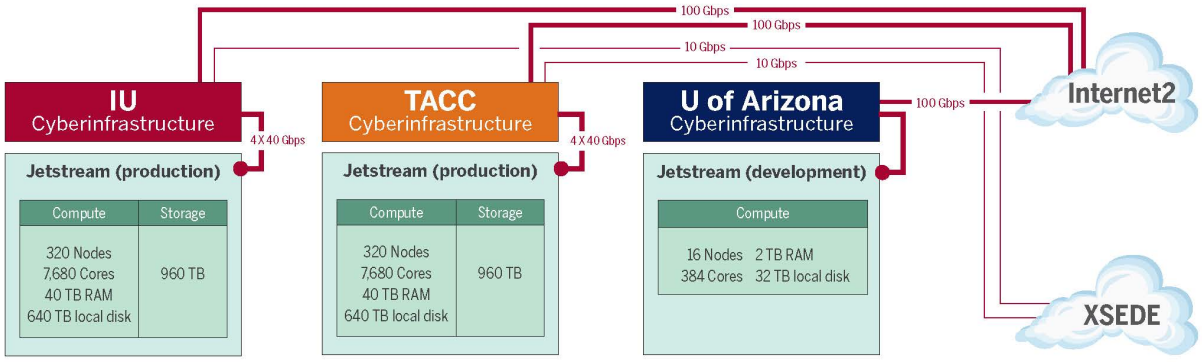


Figure 1. Diagram of Jetstream system components and national network connections.

The three hardware components of Jetstream have been in their final state of assembly since December 2015. The test system (Jetstream-AZ) has already been accepted by the National Science Foundation. An acceptance report for this system is available [27]. The two production

components are named Jetstream-IU (located in Bloomington, IN, at Indiana University) and Jetstream-TACC (housed at the Texas Advanced Computing Center on the campuses of the University of Texas at Austin).

The Jetstream-IU and Jetstream-TACC Dell PowerEdge (PE) system components were requisitioned by Indiana University for eventual delivery to the IUPTI and Texas Advanced Computing Center (TACC). This system was ordered on 07/29/2015 via the purchase order numbers 1681608 and 1681609 respectively. The respective clusters arrived at TACC's data center on 10/16/2015 and the IUB Data Center on 10/19/2015. Each of these system components has the following basic characteristics:

- Compute nodes: 320 Dell M630 blades with a total of 640 CPUs, 15,360 processor cores, 258 TFLOPS peak processing capability, and 40 TB RAM. Each blade contains two Intel "Haswell" E5-2680v3 (12-core) 2.5 GHz processors for a total of 24 processing cores resulting in a peak performance of 806.4 GFLOPS and 64 GB RAM for management and storage servers and 128 GB RAM for compute servers.
- Storage nodes: 20 Dell R730 servers, with a total of 40 CPUs, 960 processing cores, 1.2TB RAM, 16 TB local storage, 960 TB of storage and a peak processing capability of 16.1 TFLOPS.
- Management nodes: 7 Dell R630 servers, with a total of 14 CPUs, 168 processing cores, 448 GB RAM, 5.6TB local storage, and a peak processing capability of 5.6TF
- The hardware infrastructure is based upon Dell PowerEdge servers with a 10/40 Gbps Fat-Tree Ethernet fabric [28].

The cloud infrastructure is based upon OpenStack [29] with its ability to deliver virtualized compute capacity. The Atmosphere orchestration layer and interface provides the software interface directly experienced by the user. Globus authentication services manage authentication of users against the XSEDE user database. A detailed timeline of the implementation of the components of the Jetstream system is presented in Table 20.

2.11. System as purchased matches system as specified in revised statement of work

The system as purchased matches the system specified in the revised statement of work (revision based on the difference between the originally proposed budget and the final budget) the components of which are listed in Table 2.

Table 2. Hardware specifications for the Jetstream systems.

Jetstream component	# CPUs	# Cores	PFLOPS	Total RAM (GB)	Secondary storage (TB)	Node local storage (TB)	Connection to Internet2 (Gbps)	Outcome
Production components of Jetstream								
IU	640	7,680	0.258	40,960	960	640	100	G
TACC	640	7,680	0.258	40,960	960	640	100	G
Jetstream test and build system								
Arizona	32	384	0.013	2,048	192	32	100	G
Total	1,312	15,744	0.529	83,968	2,112	1,312	300	G

3. Jetstream is integrated with XSEDE

As specified in NSF solicitation 14-536 and in the Project Execution Plan, Jetstream is integrated with XSEDE.
Table 3 presents the basic criteria for integration with XSEDE specified in the solicitation NSF 14-536 and the Cooperative Agreement between NSF and IU for award 1445604. All these criteria have been satisfied.

Table 3. Basic criteria for any XD Program resource to be considered integrated with XSEDE

Criterion	Met by	Date achieved	Status
Physical network connection to XSEDE	Successful connection of Jetstream-IU and Jetstream-TACC with XSEDE system to XSEDE network	February 23, 2016	G
System available for allocation via standard XSEDE processes	Jetstream included in XSEDE allocations listings and available for users to request	September 15, 2015	G
Account management interoperability	Accounts created via receipt of packets from XSEDE AMIE system	February 10, 2016	G
Participation in the XSEDE Service Provider Forum	Jetstream admitted as a Level 1 Service Provider to the XSEDE Federation and SP Forum.	December 7, 2015	G

4. Jetstream meets the hardware performance criteria defined in the Project Execution Plan

A detailed description of the Jetstream Indiana and TACC subsystems is presented in Appendix II. The acceptance tests included in the PEP are presented in detail in Appendix III. The test methodology used in executing these tests and detailed results are presented in Appendix IV. Table 4 presents, in summary form, the results of the basic hardware performance tests specified in the PEP. All tests were passed successfully. Jetstream, as implemented, fulfills the basic hardware and capacity tests described in the PEP.

Table 4. Summary of hardware performance tests on Jetstream.

Test	Success criteria	Key test metric result achieved	Outcome
Single-Node Performance Tests			
High-Performance Linpack (HPL): Single node Linpack performance within a VM will achieve 80% of the peak floating-point performance of HPL running in the native Linux OS for a problem size that uses at least half of the on-node memory. (Measurements will be rounded to nearest %).	Achieved in VM 80% or more of in Native	Achieved floating point performance in Linux OS: 697 GFLOPS at IU and 701 GFLOPS at TACC Achieved floating point performance inside VM: 678 GFLOPS 87% on both clusters	
STREAM: Single node OpenMP threaded STREAM performance will be at least 65 GB/s (aggregate across the node). (Measurements will be rounded to nearest 1 GB/s)).	65 GB/s	100 GB/s on the Indiana cluster and 113 GB/s on the TACC cluster	
10 Gigabit Ethernet Bandwidth: the 10 GigE interface on each node will achieve at least 1 GB/s for large-message point-to-point transfers (Measurements will be rounded to the nearest 0.1 GB/s)	1 GB/s	1.1 GB/s on the Indiana Cluster and 1.2 GB/s on the TACC cluster	
File System and Storage Benchmarks			
The system will achieve a minimum of 200 MB/s data transfer rate for data reads and a minimum of 100 MB/s writes from within a virtual machine to the block storage. (Measurements will be rounded to the nearest MB/s.	200 MB/s read 100 MB/s write	244 MB/s 359 MB/s	
System Reliability Test			
During System early operations mode the system will be operated with uptime of at least 95% for a period of 14 days.	95% uptime for 14 days	100% uptime for 14 days at each site and as an integrated resource	
System Capacity Test			
Jetstream will support at least 640 VMs simultaneously	At least 640 VMs simultaneously	Jetstream-IU: 998 VMs ran simultaneously on 4/15/16. Jetstream-TACC: 832 VMs ran simultaneously on 3/7/16. Combined 1217 VMs on 4/29/16	

We note that there are significantly different test results reported for the file system and storage benchmark tests. During 2016 the finalization of a purchasing contract with Dell Inc. and shipment of production hardware by Dell Inc. was held up for approximately five months as the Program Execution Plan was sent out for peer review, and the PEP was finalized by the NSF Division of Grants and Awards. The value added to the overall project as a result of the newly

included step of peer review of the PEP (not done for awards prior to the awards for NSF 14-536) was significant. However, it created a situation in which the Jetstream team was rushing to get the system into early operations mode as quickly as possible. Therefore our testing protocol was quite simply "test till the first passing result, then quit." The apparent disparity between test results for file systems test between Jetstream-Indiana and Jetstream-TACC is a result of this approach, and the TACC test was passed during an earlier state of software configuration than the Jetstream-Indiana tests. We will re-run the tests with the system prior to finalizing a paper characterizing the performance of Jetstream, which we expect to present at the XSEDE16 conference in July of 2016.

The VM loads run on 4/15/2016 and 3/7/16 were generated synthetically (programmatically) using tiny VMs and executables that use modest amounts of CPU. These tests were performed via the Atmosphere API and via OpenStack Rally.

- Loads generated via the Atmosphere API have been designed to simulate the behavior of and experience of real users by implementing an agent-based model for interacting with the system. Under this model, 25 software agents interact with Atmosphere, launching VMs, creating and mounting block volumes, and accessing the VMs via SSH. These agents also undertake occasional management actions such as suspending, resuming, restarting, and deleting instances. Instance launches were assigned randomly between TACC and IU clouds, with a random uniform distribution of VM sizes, and a random uniform choice of featured base image. We now know that the sizes of user-initiated instance launches don't follow a uniform distribution on Jetstream (and probably not on CyVerse) but this particular load test was designed early in Jetstream's operations when we did not have a distribution on which to model an alternative distribution. We elected to not perform workloads on the launched VMs because we believed that real-world testing by users would give better data on performance of specific codes.
- Tests using OpenStack Rally have been targeted at understanding the robustness of the OpenStack control plane. The placement algorithm for Jetstream is least loaded first starting with memory, then allocated vCPU count. This is the default algorithm for OpenStack. The synthetic load testing performed demonstrates that the control plane of the system is unlikely to experience faults during normal operation. The Jetstream storage environment uses copy on write cloning and thin provisioning which minimize the impact of launching different VM sizes. The median boot time for starting 10 m1.large instances is 8.562 sec vs. 8.727 sec for 10 m1.tiny instances. Tiny instances take longer to schedule because more hypervisors have sufficient resources to start this size of instance and must be considered by the scheduler

The test on 4/29/2016 was an excellent test that represented a real operating scenario for Jetstream with a significant computational load. The 1,217 VMs running on Jetstream were a mix of jobs generated by individual people (a minority) and by the SEAGrid science gateway (the vast majority). The VMs initiated by SEAGrid were running a real docking scenario of ligands against a protein database, using the software package Dock. A mix of VM sizes was used, although most were in the smallest two size classes of VMs defined for Jetstream. CPU load peaked at over 20%. The VMs were run across both production elements of Jetstream simultaneously and real scientific work of value to researchers who make regular use of SEAGrid.

4.1. Acceptance test criteria and results: software-delivered capabilities

The PEP for Jetstream describes a number of capabilities that Jetstream is to deliver to users who are authorized to use the system. Capability tests are demonstrated by showing that something can be done, not with tables of numbers. During the review meetings, we will show these capabilities in demonstrations. For the purposes of creating a documentation record, we include here screen shots of critical steps in the demonstrations of these capabilities.

4.1.1. Academic self-serve cloud services

The text of this acceptance test in the PEP states:

> *Provide "self-serve" academic cloud services, enabling researchers or students to select a VM image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing. Authentication to this "self-serve" environment will be via Globus.*

Implicit in the sense of the words 'cloud services' is that the two production components of Jetstream function as parts of an integrated whole. There are both capability and capacity issues to providing a cloud environment.

While not the most formal test ever conducted, Figure 2 shows a screen shot of a tweet that addresses this test.

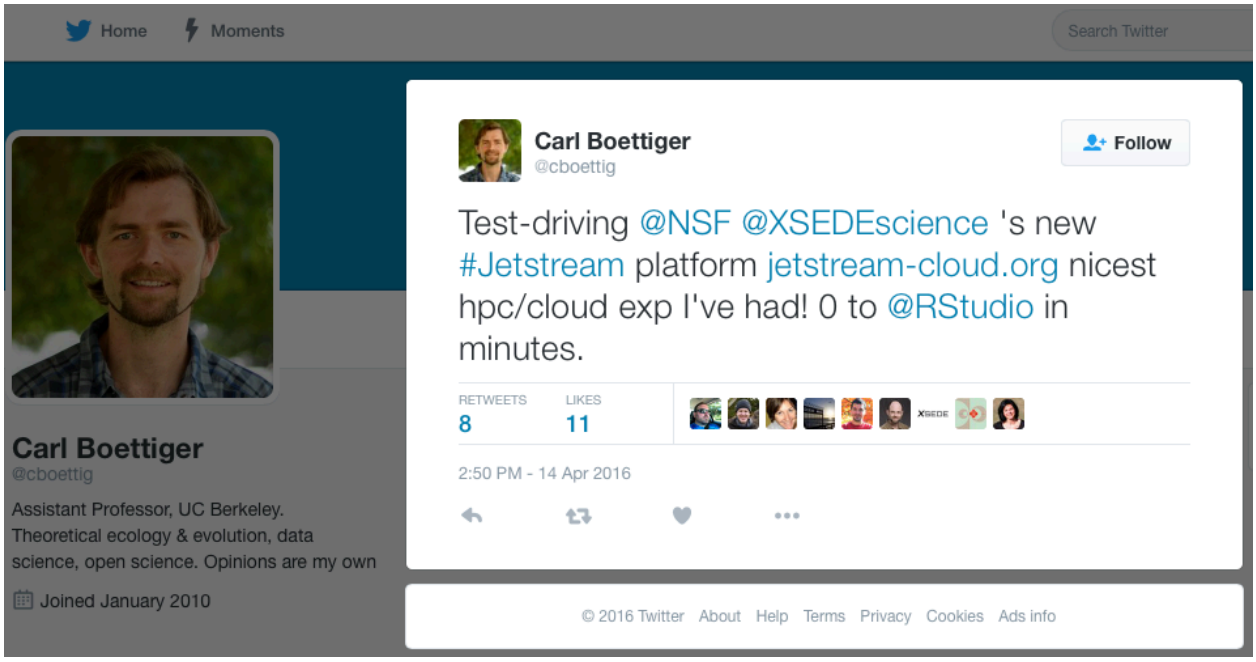


Figure 2. Tweet on April 14, 2016 from Dr. Carl Boettiger (@cboettig), Assistant Professor of Biology, UC Berkeley

Table 5 gives the steps defined in the PEP that constitute success in demonstrating the capabilities included as part of the "self-serve" academic cloud services, the status of those capabilities, and includes links to a set of Figures which are screen shots demonstrating these steps taking place.

Table 5. Acceptance test results for Jetstream as an academic, self-serve cloud service

Capability	Link to screenshot	Status
An authorized and knowledgeable user will be able to authenticate to the Jetstream user interface (which uses Globus as the mechanism for verification of credentials).	Figure 3	G
After so doing, an authorized and knowledgeable user will be able to launch a virtual machine from a menu of pre-packaged VMs on the production hardware located in Indiana or Texas.	Figure 4	G
After so doing, an authorized and knowledgeable user will be able to quiesce a VM image running on production hardware in Indiana or Texas, move it from one production system to another, and reactivate said VM.	Figure 5	G
An authorized and knowledgeable user can create and access persistent cloud storage on the Indiana or Texas production hardware	Figure 6	G
An authorized and knowledgeable user can modify a preexisting VM image and manually store that VM image to one of the production locations within Jetstream.	Figure 7	G

Figure 3 shows the Jetstream Atmosphere web interface uses Globus as the mechanism for verification of credentials (documentation for users in https://iujetstream.atlassian.net/wiki/display/JWT/System+Access).

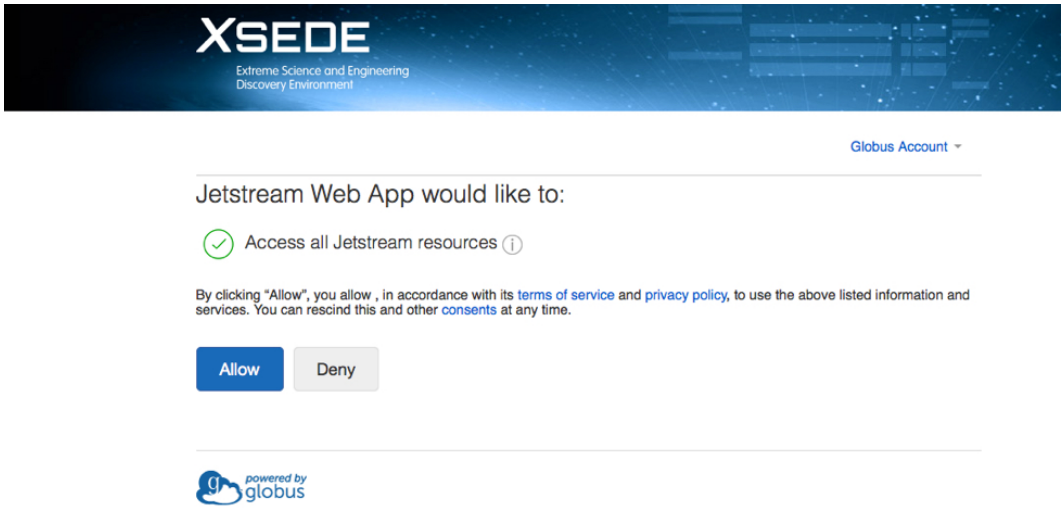


Figure 3. Screenshot demonstrating the Jetstream user authentication interface.

Once logged in, the user selects an image from the listing of "Featured Images" and then can launch it on either Jetstream cloud. For a screenshot, see Figure 4; also, user documentation available at: https://iujetstream.atlassian.net/wiki/display/JWT/Launching+your+VM

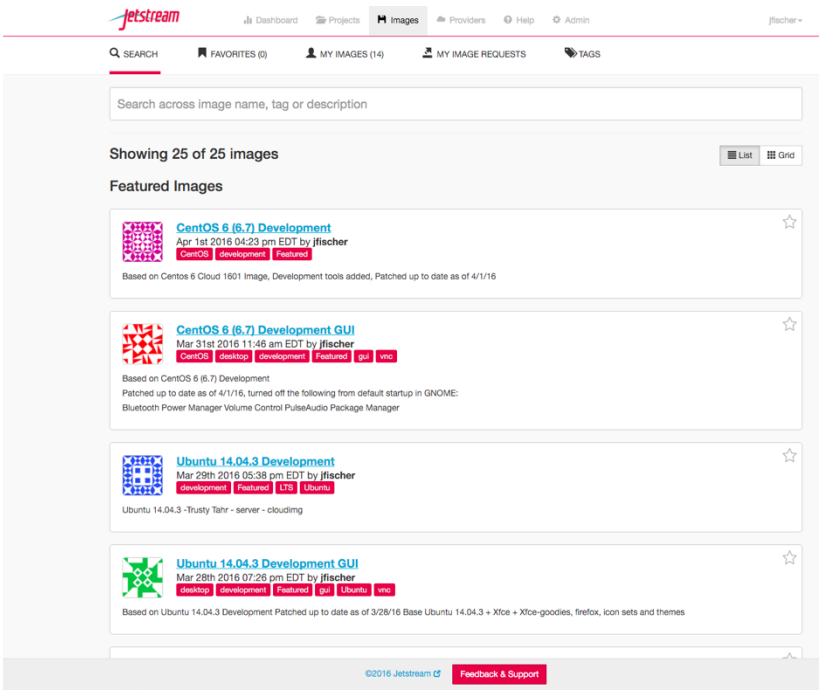


Figure 4. Screenshot demonstrating launching of a VM image.

Featured images are VM images provided to every user of Jetstream and certified by the Jetstream team to function properly. A full list of featured images currently available is presented in Table 14. The featured images include a number of basic Linux OS installations which can be used as a starting point for creating new images and also include specific scientific functions desired by end users.

Users may launch an instance on one component of the production Jetstream system (TACC or Indiana), quiesce it, move it to the other component of the production system, and activate it there. The instance in Figure 4 shows part of this process. The instance was initially launched on the IU cloud, stopped, a snapshot was created, it was migrated to the TACC cloud, and re-launched. Users may also create volumes for persistent storage on Jetstream. They create volumes up to their storage allocation limit to use on any VM instances they create (see Figure 5). Figure 6 shows access of storage from within a VM. This process is documented completely at https://iujetstream.atlassian.net/wiki/display/JWT/Customizing+and+saving+a+VM

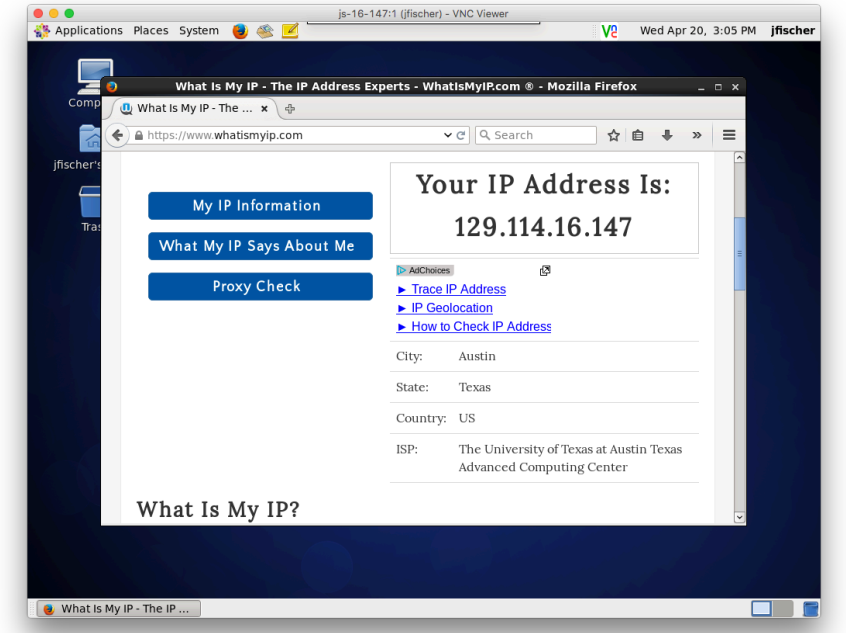


Figure 5. Screen shot showing an intermediate step in creating, quiescing, and moving an image between IU and TACC subcomponents of Jetstream.

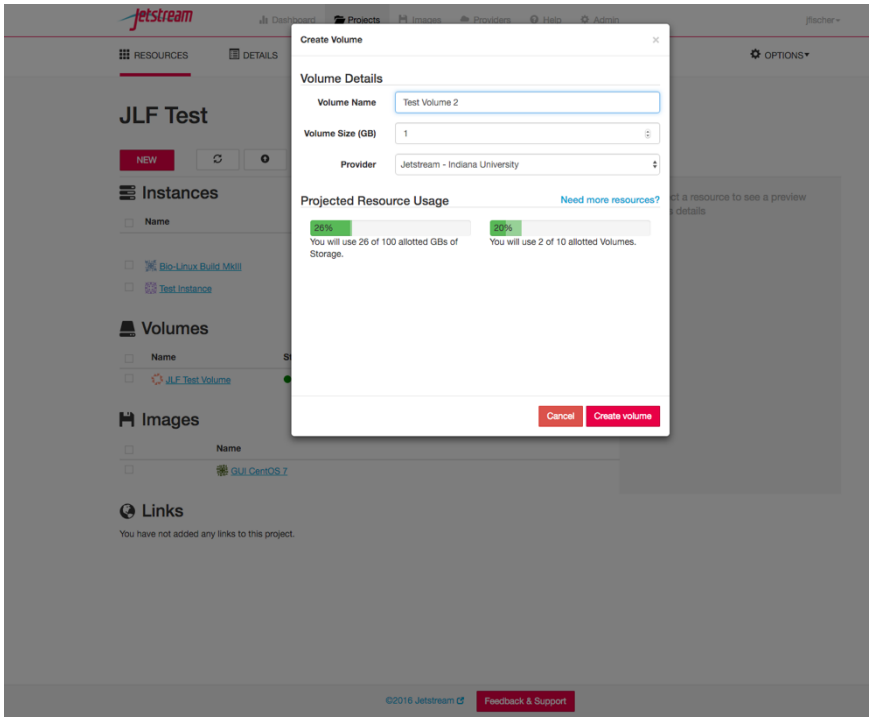


Figure 6. Screen shot showing access of storage.

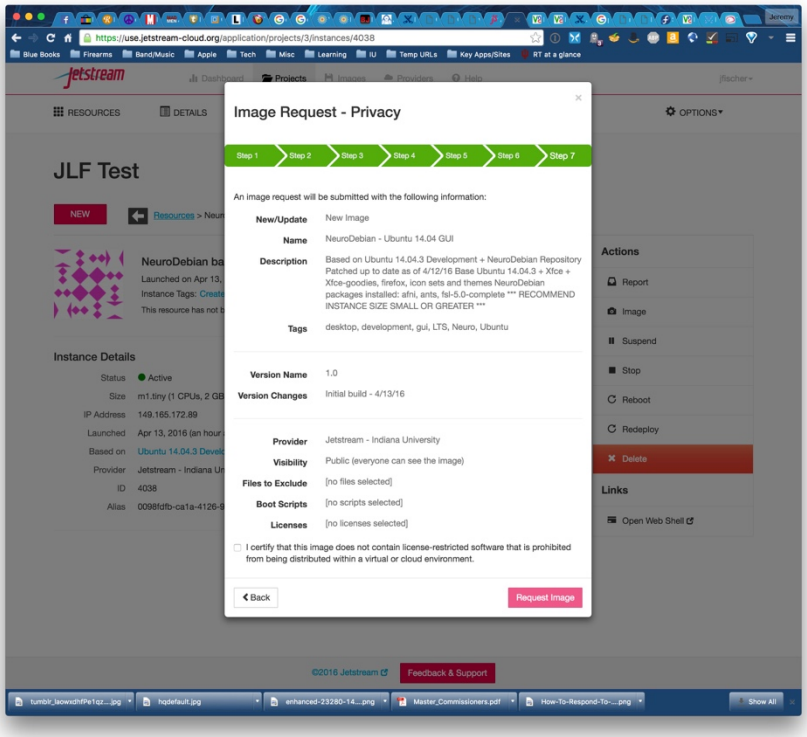


Figure 7. Screen shot showing storage of a modified VM

4.1.2. *Host Persistent Science Gateways*

The text of this acceptance test in the PEP states:

> *Jetstream will support persistent science gateways, including the capability of hosting persistent science gateways within a VM when the nature of the gateway is consistent with operation within a VM. Galaxy will be one of the initial science gateways supported.*

These criteria are also quantitative, and presented in Figure 8. Detailed results and test output are included in Appendix V. Jetstream as implemented successfully fulfills all of the tests and metrics defined in the PEP. Figure 8 is a screen shot of SEAGrid running on Jetstream.

Table 6 shows success criteria.

Table 6. Acceptance test results for Jetstream as a host of persistent science gateways.

Test	Success criteria	Key test metric result achieved	Outcome
Galaxy gateway availability and correct function.	The Galaxy bioinformatics gateway is installed and will operate a demonstration workflow providing correct results, based on comparison with output results from a known reference installation. The job will complete within 25% of the time required to complete an analysis running on equivalent[3].	Galaxy was installed in production mode on 4/15/2016. Galaxy on Jetstream produces correct results on a test workflow suggested by the Galaxy PI Dr. James Taylor Execution on Jetstream is 41% of execution time on Mason, (56% if normalized by clock speed) and 80% of the time required when run on Stampede	G
One other exemplar science gateway that is known to function properly in other XSEDE-supported gateway hosting environments will function on Jetstream.	The gateway will function and remain reliable to within 2% of the overall system availability achieved during system reliability tests during a 14-day test period (e.g. if the system turns out to be available with an uptime of 96%, the gateway used to test this criterion will be available 96% - 2% or 94%).	Two SciGAP gateways began operating on 4/15/2016: PGA generic portal: http://js-172-125.jetstream-cloud.org/; SEAGrid: http://js-172-132.jetstream-cloud.org/ These gateways have operated continuously for >14 days, from 4/15/16 to the writing of this report.	G

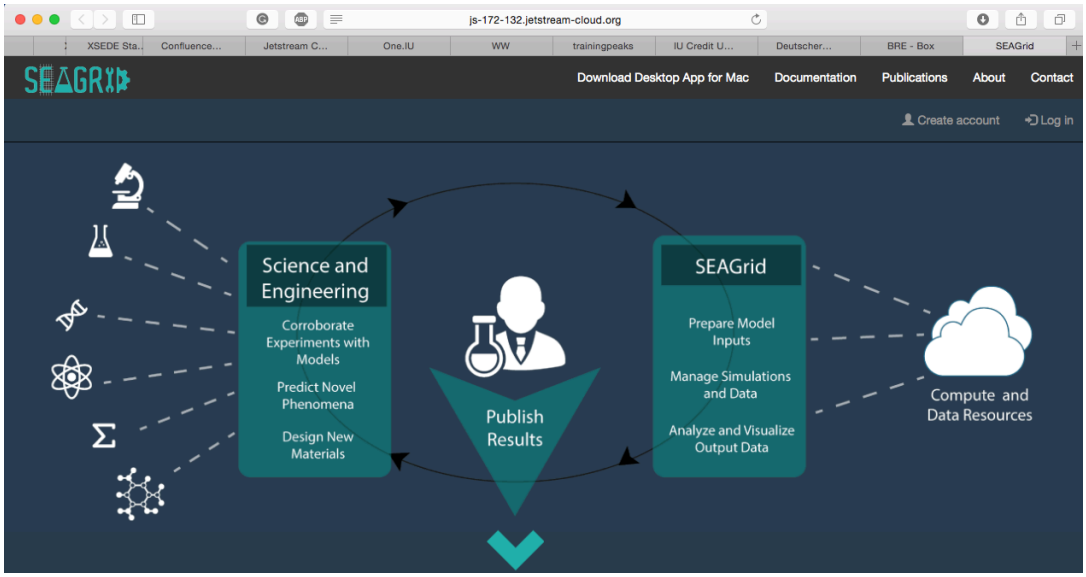


Figure 8. A screenshot of SEAGrid operating on Jetstream.

[3] By "within 25%" what was intended is the Jetstream execution time would not be more than 25% slower than the time required to run on an equivalent system. We did not count on the execution being significantly faster on Jetstream than other systems when this criterion was written.

4.1.3. *Data movement, storage and dissemination*

The PEP describes these capabilities as:

- *Jetstream will support data transfer with Globus Connect.*
- *Users will be able to store VMs in the Indiana University persistent digital repository, IUScholarWorks (scholarworks.iu.edu) and obtain a Digital Object Identifier (DOI) that is associated with the VM stored.*

The performance characteristics of the storage system are verified through the file system tests presented earlier. Globus Connect is a service offered by a partner organization that contains a set of performance characteristics that are well understood, and not affected by this solicitation. As stated in the PEP, the first item above becomes a functionality test:

- *An authorized and knowledgeable user can select a file to which they have rights on a system outside Jetstream, and move that file and save it on storage on Jetstream (with the condition that the file size is within the storage quota set for their use on Jetstream).*
- *An authorized and knowledgeable user can select a file to which they have rights on Jetstream, and move that file and save it on storage to a system on which that user has rights and which is accessible from open public networks (with the condition that the file size is within the storage quote set for their use on Jetstream).*

The second feature described above is again a capability test, satisfied by the following:

- *An authorized and knowledgeable user can successfully save a VM previously stored to disk storage on Jetstream into a format supported by DSpace, upload that file to IU Scholarworks.iu.edu, and using the existing online forms submit that document for publication via IUScholarWorks. Subsequent to that, provided the relevant and required information has been provided by the user, the VM will appear in IUScholarWorks and the user will receive a DOI identifier for that object. Note: This is a "human in the loop" process and may take days from upload and submission to publication and receipt of DOI. Email transactions may be required beyond the initial submission.*

All of the capabilities stated in the PEP may now be achieved successfully as shown below in

Table 7. As before, this table includes links to screen shots of these activities underway. These capabilities will be demonstrated during the acceptance review meetings.

Table 7. Acceptance test results for Jetstream with regards to data movement and dissemination

Capability	Link to screen shot or additional detailed results	Status
An authorized and knowledgeable user can select a file to which they have rights on a system outside Jetstream, and move that file and save it on storage on Jetstream (with the condition that the file size is within the storage quota set for their use on Jetstream).	Figure 9	
An authorized and knowledgeable user can select a file to which they have rights on Jetstream, and move that file and save it on storage to a system on which that user has rights and which is accessible from open public networks (with the condition that the file size is within the storage quote set for their use on Jetstream).	Figure 10	
An authorized and knowledgeable user can successfully save a VM previously stored to disk storage on Jetstream into a format supported by DSpace, upload that file to IU Scholarworks.iu.edu, and using the existing online forms submit that document for publication via IUScholarWorks. Subsequent to that, provided the relevant and required information has been provided by the user, the VM will appear in IUScholarWorks and the user will receive a DOI identifier for that object. Note: This is a "human in the loop" process and may take days from upload and submission to publication and receipt of DOI. Email transactions may be required beyond the initial submission.	Figure 11 & Figure 12	

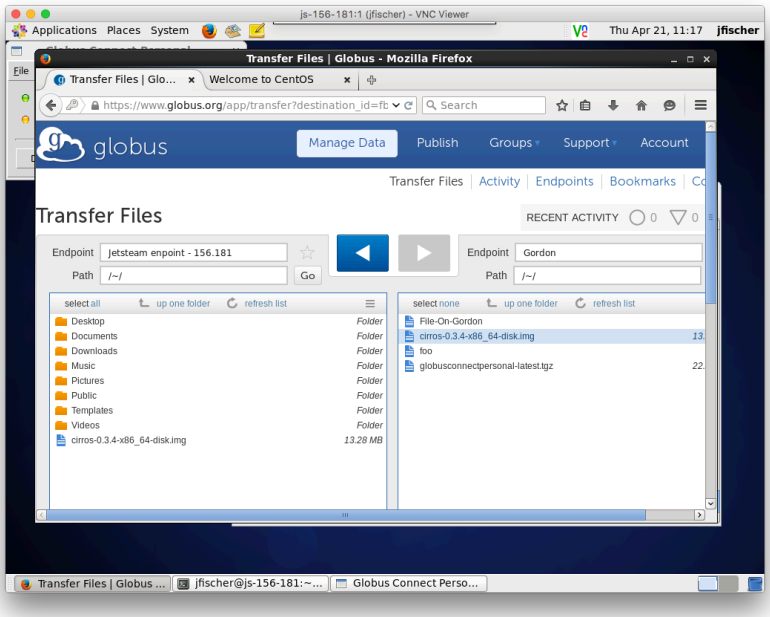


Figure 9. A user moving a file from Gordon and saving it on storage on Jetstream

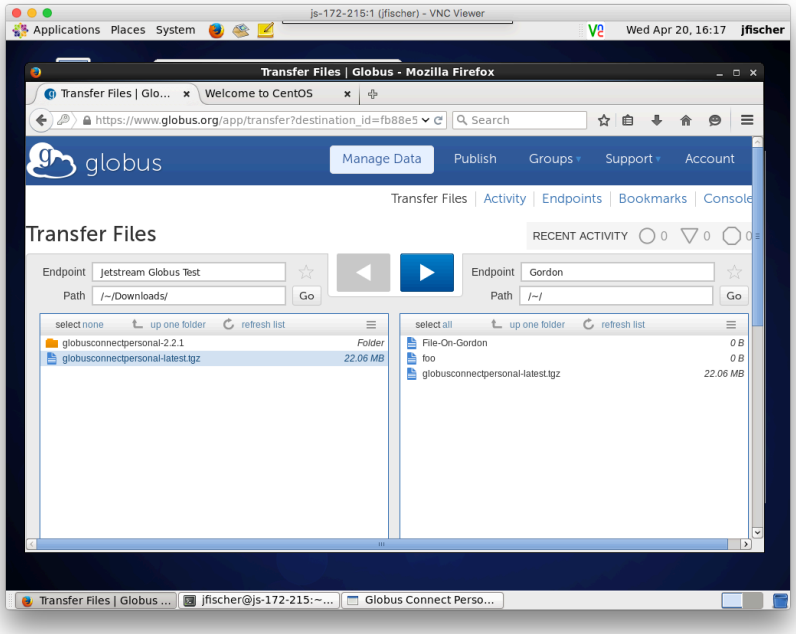


Figure 10. A user moving a file from Jetstream and saving it on Gordon.

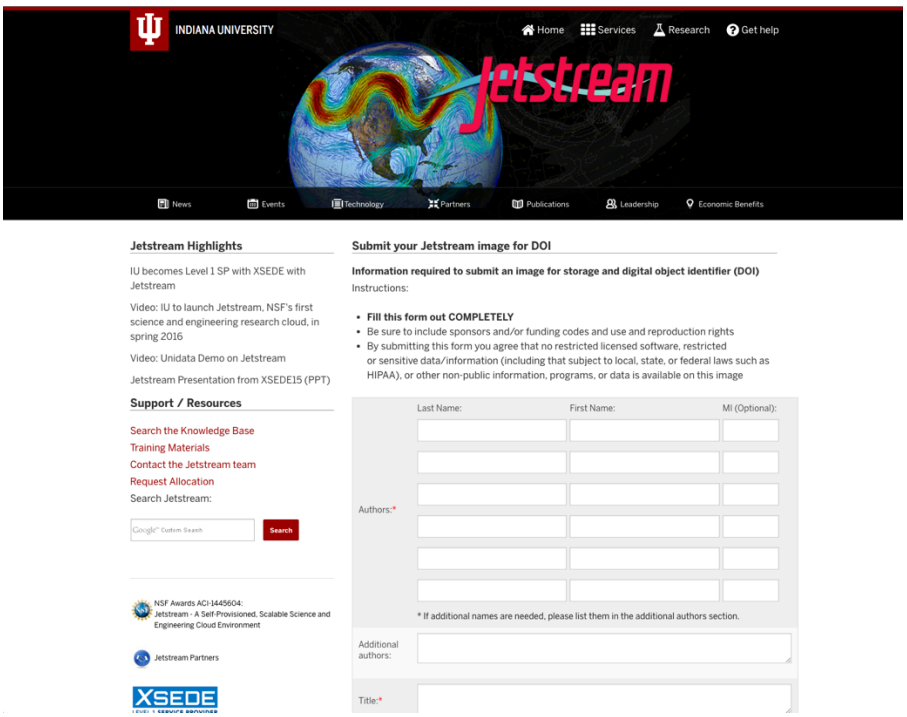


Figure 11. A user filling out the required form providing minimal Dublin Core information in order to upload a VM to the IUScholarWorks persistent digital repository

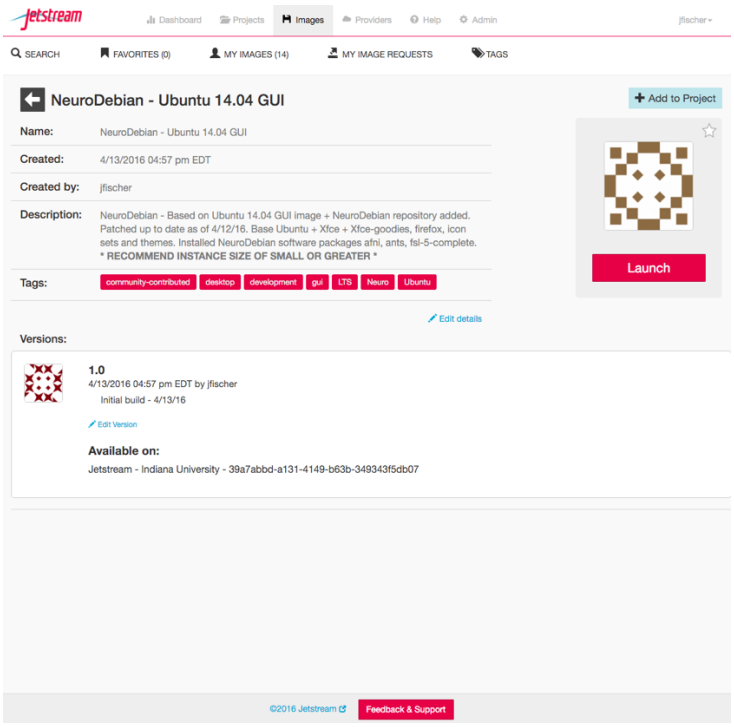


Figure 12. A VM newly made available for download (and subsequently use) by anyone with a network connection and web browser.

4.1.4. *Provide virtual Linux desktop services delivered from Jetstream to tablet devices*

The full text of this capability as described in the PEP is 'Provide virtual Linux desktop services delivered from Jetstream to tablet devices. This service is aimed to increase access to Jetstream for users at institutions with limited resources including small schools, schools in EPSCoR states, and Minority Serving Institutions.'

This test is a functionality test, with some time constraints. According to the PEP, this test is satisfied by the following:

> *An authorized and knowledgeable user can access Jetstream from a tablet device, and load a virtual Linux desktop configured in a way that allows the user to access Jetstream services.*

This test is satisfied, as summarized in Table 8 and subsequent screenshot (Figure 13).

Table 8. Acceptance test results for Jetstream with regards to dissemination

Capability	Link to screen shot or additional detailed results	Status
An authorized and knowledgeable user can access Jetstream from a tablet device, and load a virtual Linux desktop configured in a way that allows the user to access Jetstream services	Figure 13	G

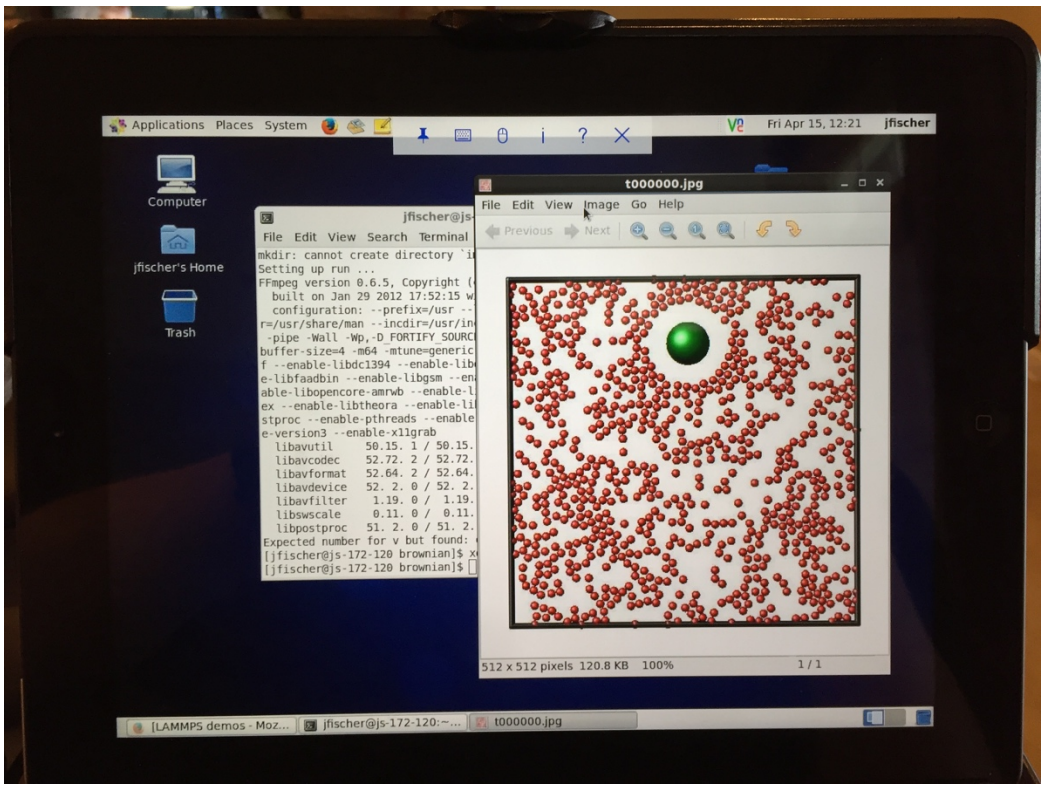


Figure 13. Accessing a Jetstream virtual desktop from a 2nd generation iPad via cellular network from Kentucky.

4.2. Summary of PEP-specified acceptance test results

The simple summary of all of the performance test results above, in our view, is that Jetstream as implemented now is indeed the system proposed in our final proposal documents including the revised statement of scope of work based on the reduction in budget between what we proposed and the final awarded budget. All of the acceptance test criteria specified in the Project Execution Plan have been completed successfully.

5. Jetstream is allocatable and allocated at 90% capacity

Beyond passing the system acceptance tests as proposed in the PEP, in the process of implementing a first-of-a-kind CI resource it is useful to document tangible evidence of the utility of the system to the national open science research community. In this section we discuss success in meeting one of the NSF-specified criteria – that the system be allocated at 90% capacity to the national research community. This criterion admits two interpretations: the organization operating the resource will make 90% of the resource available for allocation. The other interpretation is that the XSEDE Resource Allocation Committee (XRAC) and XSEDE

staff will authorize allocations that constitute 90% of the available resource. Jetstream at present satisfies both interpretations.

5.1. Available to allocate at 90% of capacity

Jetstream SUs (XSEDE Standard Units) are defined as service units (see Table 9).

Table 9. Definition of SUs

VM Size	vCPUs	RAM (GB)	Local Storage (GB)	SU cost per hour
Tiny	1	2	8	1
Small	2	4	20	2
Medium	6	16	60	6
Large	10	30	120	10
XLarge	22	60	240	22
XX Large	44	120	480	44

IU authorized XSEDE to begin allocating Jetstream resources via its standard allocation processes as of September 15, 2015. It is both 90% available for allocation by XSEDE and 90% allocated. Details follow below.

Table 10. Jetstream is 90% available for allocation and 90% allocated for the next several months

Allocation metric	Status
Jetstream is made available at 90 % of its total production capacity to XSEDE for allocation	G
Jetstream is allocated by XSEDE at 90% of its total production capability by XSEDE	G

In a 30-day month, Jetstream produces 1.67 million SUs at 96% uptime (the required level of uptime specified by the NSF solicitation 14-536). In a quarterly allocation process, 90% of the cycles available on XSEDE would constitute 4.5 M SUs per quarterly allocation cycle. Initially, we made modest amounts of Jetstream allocatable via XSEDE mechanisms starting with 2.5M SUs in the October 2015 round of allocations. This was done on the advice of XSEDE allocation and accounting staff. This advice was offered by XSEDE staff to avoid a situation in which many millions of SUs would be awarded before the system is officially accepted and in production. The Jetstream team has authorized XSEDE to allocate 5M SUs per quarterly cycle of allocations beginning the allocation request cycle that closes on July 15, 2016. As of the writing of this report, then, Jetstream fulfills the criterion of having 90% of the resource available for allocation by XSEDE.

Allocations to date for XSEDE total up to 8,128,060 SUs (see Table 11). If Jetstream goes into production operations on June 1, 2016, current allocations can use Jetstream's full capacity from then till late October 2016. Between now and then there will be two additional cycles of XRAC (XSEDE Resource Allocation Committee) allocation awards, plus ongoing awards for startup

allocations on Jetstream. Two additional cycles of XRAC allocations should extend the period in which Jetstream is 90% allocated by several months in the future.

Jetstream is allocated at 90% capacity for the next several months and we expect it to remain so for the foreseeable future. There are two requests pending before the XRAC in the current round of resource requests, to be reviewed in June, that total more than 800,000 SUs.

Table 11. Allocations of resources on Jetstream as of April 25, 2016.

PI	PI Institution	Field	On behalf of national consortium or work at home institution	Allocation Type	SUs	Home State of PI	EPSCoR State	MSI, HSI, Tribal College, or HBCU*?
Gopu, Arvind	WIYN / IU	Astronomy	Nat'l	Startup	250,000	IN		
Kremin, Anthony	U. Michigan	Astronomy	Home	Startup	50,000	MI		
Doak, Thomas	NCGAS / IU	Bio	Nat'l	Supplement	250,000	IN		
Zimmerman, Naupaka	University of Arizona	Bio	Home	Startup	50,000	AZ		
Hill, Joshua	Texas A&M University	Bio	Home	Startup	100,000	TX		
Watson, Deborah	Bloomsburg University of Pennsylvania	Bio	Home	Startup	50,000	PA		
Mutangadura, Tendai	University of Missouri, Columbia	Bio	Home	Startup	150,000	MO	Yes	
Merchant, Nirav	iPlant / University of Arizona	Bio	Nat'l	XRAC	1,000,000	AZ		
Taylor, James	Johns Hopkins University	Bio	Nat'l	XRAC	500,000	MD		
Brendel, Volker	IU	Bio	Nat'l	PI Discretion	250,000	IN		
Buechlein, Aaron	IU	Bio	Home	Startup	50,000	IN		
Boettiger, Carl	University of California, Berkeley	Bio (Field Biology)	Home	Startup	50,000	CA		
Culich, Aaron	University of California, Berkeley	Campus Champion	Home	Training	50,000	CA		
Harvey, Russ	University of California, Riverside	Campus Champion	Home	Training	50,000	CA		
Gazula, Vikram	University of Kentucky	Campus Champion	Home	Training	50,000	KY	Yes	

PI	PI Institution	Field	On behalf of national consortium or work at home institution	Allocation Type	SUs	Home State of PI	EPSCoR State	MSI, HSI, Tribal College, or HBCU*?
Marinshaw, Ruth	Stanford University	Campus Champion	Home	Startup	50,000	CA		
Nickel, Ben	Idaho National Lab	Campus Champion	Fed Gov't	Startup	50,000	ID	Yes	
Smith, Jack	Marshall University	Campus Champion	Home	Startup	10,000	WV	Yes	
Basheer, Ershaad	Temple University	Campus Champion	Home	Startup	50,000	PA		
Brooks, Emre	UT San Antonio	Chemistry	Home	Startup	17,250	TX		Yes
Beck, Brian	UTA/ TACC	Computational Science	Home	Startup	250,000	TX		
Skow, Dane	University of Wyoming	Computational science	Home	Startup	100,000	WY	Yes	
Jha, Shantenu	Rutgers University	Computational Science	Nat'l	XRAC	200,000	NJ		
Pierce, Marlon	IU, SciGAP	Computational science (gateways)	Nat'l	Supplement	50,000	IN		
Pummil, Jeff	University of Arkansas	Computational Science, Bio	Home	Startup	100,000	AR	Yes	
von Laszewski, Gregor	IU	Computer science	Home	Renewal	50,000	IN		
VanRenesse, Robert	Cornell	Computer Science	Home	Startup	50,000	NY		
Rudolph, George	The Citadel	Computer Science	Home	Startup	50,000	SC		
Wong, Kwai L.	University of Tennessee	Computer Science	Home	Startup	50,000	TN	Yes	
Hicks, John	Internet2	Computer Science	Nat'l	Education	500,000	IN		
McCaulay, Scott	IU	Computer Science	Home	Startup	100,000	IN		
Rajamohan, Srijith	Virginia Polytechnic Institute and State University	Engineering	Home	Startup	50,000	VA		
Boettiger, Carl	University of	Bio (Field	Home	Startup	50,000	CA		

PI	PI Institution	Field	On behalf of national consortium or work at home institution	Allocation Type	SUs	Home State of PI	EPSCoR State	MSI, HSI, Tribal College, or HBCU*?
	California, Berkeley	Biology)						
Falgout, Jeff	US Geological Survey	Geo	Fed Gov't	Startup	50,000	CO		
Daniels, Michael	National Center for Atmospheric Research	Geo	Nat'l	Startup	50,000	CO		
Fils, Douglas	Consortium for Ocean Leadership	Geo	Nat'l	Supplement	50,000	DC		
Graves, Sara	University of Alabama Huntsville	Geo	Nat'l	Startup	100,000	AL	Yes	
Ramamurthy, Mohan	UCAR / Unidata	Geo	Nat'l	Startup	50,000	CO		
Ahern, Timothy	IRIS / U. Washington	Geo	Nat'l	Startup	105,120	WA		
Phillips, James	University of Illinois at Urbana-Champaign	Molecular biosciences	Home	Startup	50,000	IL		
Nagy, Laszlo	Sanford-Burnham Medical Research Institute	Molecular biosciences	Home	Startup	100,000	CA		
Ma, Lijun	University of California, San Francisco	Molecular biosciences	Home	Startup	50,000	CA		
Zhang, Yang	University of Michigan	Molecular biosciences	Home	Startup	50,000	MI		
Reddy, Karen	Johns Hopkins University School of Medicine	Molecular biosciences	Home	Startup	100,000	MD		
Freeberg, Mallory	Johns Hopkins University School of Medicine	Molecular biosciences	Home	Startup	100,000	MD		
Rossi, Miriam	Vassar	Molecular	Home	XRAC	275,000	NY		

PI	PI Institution	Field	On behalf of national consortium or work at home institution	Allocation Type	SUs	Home State of PI	EPSCoR State	MSI, HSI, Tribal College, or HBCU*?
	College	biosciences						
Xu, Jinbo	Toyota Technological Institute at Chicago	Molecular biosciences (Deep learning)	Home	Startup	50,000	IL		
Borner, Katy	IU	Network Science	Nat'l	Startup	50,000	IN		
Cleveland, Sean	University of Hawaii	Ocean Science	Home	Startup	100,000	HI	Yes	Yes
Tao, Jian	Louisiana State University	Ocean Science	Home	Startup	50,000	LA	Yes	
Onyisi, Peter	Atlas / University of Texas at Austin	Physics	International	XRAC	2,119,920	TX		
Fischer, Jeremy	IU	Resource Provider staff	--	Startup	250,000	IN		
Borner, Katy	IU	Visualization	Home	Education	150,000	IN		
Total	**54 Allocations**				**8,577,290**			

Demonstration of potential value to the US science and engineering research community: allocations, letters of support, availability and contribution of VMs, and number of users actually trying Jetstream

In this section, we attend to the use of Jetstream in early operations and its dual role as pilot implementation for the NSF – a first-of-a-kind production cloud system – and its potential utility as a production system for the US open science and engineering research community. In this section in particular we will present data that demonstrates that:

- Jetstream is interesting to researchers and research educators, as expressed through allocation requests, requests for letters of support, and actual use of the system
- The system is usable – it has a variety of software tools available that make it useful to the majority of the user communities identified as intended users of Jetstream.

5.2. Further analysis of allocations to date

Allocation data serve at least two purposes in this report: they are used to demonstrate fulfillment of the 90% allocated metric specified by the NSF; they also are indicators of interest. A tally of allocations approved to date is provided in Table 12, below. Of the allocations to date, one is for an international research collaboration (ATLAS), two are for US federal government research organizations, 16 are for national research collaborations (four of these 16 are for Jetstream, XSEDE, iPlant, and Galaxy). PIs with allocations are now found in 23 states and the District of Columbia. There are allocations to PIs in 10 EPSCoR states: Alabama, Arkansas, Kentucky, Hawaii, Idaho, West Virginia, Louisiana, Tennessee, Wyoming, Missouri. There are allocations so far to PIs at two Minority Serving Institutions.

In terms of the types of allocations, 1 is a renewal of an earlier allocation on a different system, 1allocation so far has been made at the PI's discretion, 38 are startup allocations, 3 are supplements to existing allocation awards, 2 are for educational purposes (university credit-bearing), 3 are for training (not credit-bearing), and 5 are large allocations (> 200,000 to > 2,000,000 SUs) made via the XSEDE Resource Allocation Committee.

Table 12. Distribution of allocations by field or area of interest.

Discipline or area of interest	Number of allocations	Number of SUs allocated
Astronomy	2	300,000
Biological sciences other than molecular biosciences	11 (including 2 for field biology)	2,500,000
Campus champions	7	310,000
Chemistry	1	17,250
Computational Science	5	700,000
Computer Science	6	800,000
Engineering	1	50,000
Geosciences	6	405,120
Molecular biosciences (protein structure, molecule docking)	8	775,000
Network Science	1	50,000
Ocean Science	2	150,000
Physics (ATLAS)	1	2,119,920
Visualization	1	150,000
XSEDE and Jetstream staff training	1	250,000

The table above demonstrates early success in one of the goals stated by the NSF in solicitation 14-536: increasing the diversity of users and uses of resources of the XD program. Relative to typical allocations on large clusters supported by XSEDE, the allocations for Jetstream show much more interest on the part of geoscientists, biologists (working in areas other than molecular biosciences), and ocean scientists. A rate of roughly 2% of allocations going to engineers is also higher than XSEDE as a whole. These are early data, but so far the interest in Jetstream as indicated by allocations suggests success in engaging disciplines and subdisciplines not traditionally highly represented among uses of other XD program resources.

5.3. Letters of commitment requested and provided

Another indication of the desirability of Jetstream as a computational resource is the number of letters of collaboration of commitment that PI Stewart has been asked to write in support of other researchers. Information on the 15 letters requested and provided so far is listed in Table 13.

Researchers obtain allocations of resources from the XD program (XSEDE and individuals SPs) by making a request via the XSEDE web portal. Modest startup accounts are approved by XSEDE staff. Larger requests go to a committee convened by XSEDE – the XSEDE Resource Allocation Committee (XRAC). Individual researchers encounter a significant learning curve when first using this process. That is, proposals that fail in some technical way may get no allocation of an XD resource even though the research is meritorious and the request is reasonable. The allocations committee also demonstrates a learning curve when dealing with a new resource; requests retrospectively seen as reasonable are sometimes rejected in the early phases of a system being available for reviews because it takes time for the XRAC to understand new resources and their capabilities. (We should note here that reviewers are volunteers and there are significant demands of time and travel to be an XRAC member).

Occasionally, principal investigators make requests for allocations of XD program resources for seemingly meritorious research just to see requests rejected outright or significantly reduced. In general, neither the XSEDE PI (John Towns) nor the PIs of any particular service providers have the ability to make a commitment on behalf of the XSEDE allocation process. Many letters of support thus say something like "we can't make any promises that you will actually be able to use system _____, but we will help you apply for resources." During the early days of Jetstream, we have taken a slightly different approach. Ten percent of SUs on Jetstream are allocable at PI discretion, per the NSF solicitation and PEP. PI Stewart has been making use of this discretionary allocation to make concrete commitments of resources in letters of commitment. That is, Stewart in a letter of commitment to PIs who want to use Jetstream writes something to the effect that we will help the PI through the XSEDE allocation process, but if that fails the Jetstream PI's discretionary time will be used to ensure that the letter recipient is assured that time on Jetstream are made available to her/him. A sample of such a letter is included as Appendix VI.

Table 13. Letters of support requested and approved on Jetstream

Home State of Principal Investigator	Number of Letters of Support Written
AZ	1
IN	7
MD	1
NC	2
ND	1
NJ	1
VA	1
WA	1
Total	**15**

5.4. Utility to disciplines of science as indicated by availability of Virtual Machines

One indication of the utility of the Jetstream system to the user community is the number of VMs available for use by the research community. We define three types of VMs:

- "Featured" VMs – VMs that the Jetstream team certifies as guaranteed to function properly, and which the Jetstream team takes responsibility for correcting if a failure to perform correctly is ever detected;
- Community Contributed VMs, which are available for use by the community and provided by someone or some group other than the Jetstream team (we make no commitment to correct failures for such VMs although providers of such may);
- and Private VMs, which are just what they sound like – private to an individual or group.

There are at present a total of 14 VMs available to users of Jetstream – 7 "featured," 7 "contributed" and each of these contributed VMs from a different group. Four of the featured VMs are already available within IU's persistent digital repository (IUScholarWorks).

The availability of these VMs speaks to the potential utility of the system to researchers and research educators. The fact that half of the VMs available at this point are contributed by groups outside of the funded Jetstream collaboration speaks to the value the community places on this sort of vehicle for doing research and collaborating on the provision of research tools.

Table 14. Featured and contributed images available via Jetstream.

Fields and functions	Featured VMs	Contributed VMs	DOIs for VMs	Contributor (when not part of funded Jetstream team)
VMs of general use to the research community				
Basic Linux OS images for developing VMs containing other applications				
	Ubuntu 14.04.3 Development		doi.org/10.5967/P9CC7T	
	Ubuntu 14.04.3 Development GUI		doi.org/10.5967/P9H59R	
	CentOS 6.7 Development		doi.org/10.5967/P97P4J	
	CentOS 6 (6.7) Development GUI			
	CentOS 7.2 Development		doi.org/10.5967/P93W2M	
File Movement				
	Wrangler iRODS -- CentOS 6.7			
Statistical analyses				
		R in Ubuntu 14.04 Docker OpenSci Project Container Build (ropensci.org)		Contributed by Boettiger Lab, UC Berkeley
VMs of use to particular disciplines, virtual organizations, or communities of practice				
Biology				
	Galaxy 16.01 Standalone			
		Genome annotation, MAKER 2.31.8 with CCTools 5.4.		
		ASTRAL – Genome scale coalescent species tree estimation, Ubuntu 14.04.3 Phylogenetics		T.Chafin, University of Arkansas, Fayetteville
		NeuroDebian - Ubuntu 14.04 GUI		Franco Pestilli, IU
		Newbler DNA genome assembly		Contributed by D. Rice, NCGAS
Computer & Network Science				
		OpenFlow network simulation and management - Operating Innovative Networks (OIN) network education VM		Contributed by John Hicks, Internet2
		Network analysis and visualization		
Totals				
	7 Featured VMs	7 Contributed VMs	5 VMs deposited in IU's persistent digital repository	5 Different contributors outside of Jetstream

5.5. Interest in use of Jetstream as demonstrated by use of Jetstream

Perhaps one of the strongest demonstrations of interest in use of Jetstream that could exist is actual use of Jetstream. There are 758 Galaxy users who have accessed Jetstream. In addition, there is growing demand from other displinnes. Figure 14 shows the increase over time of (non-Galaxy) people who have logged into Jetstream and used it.

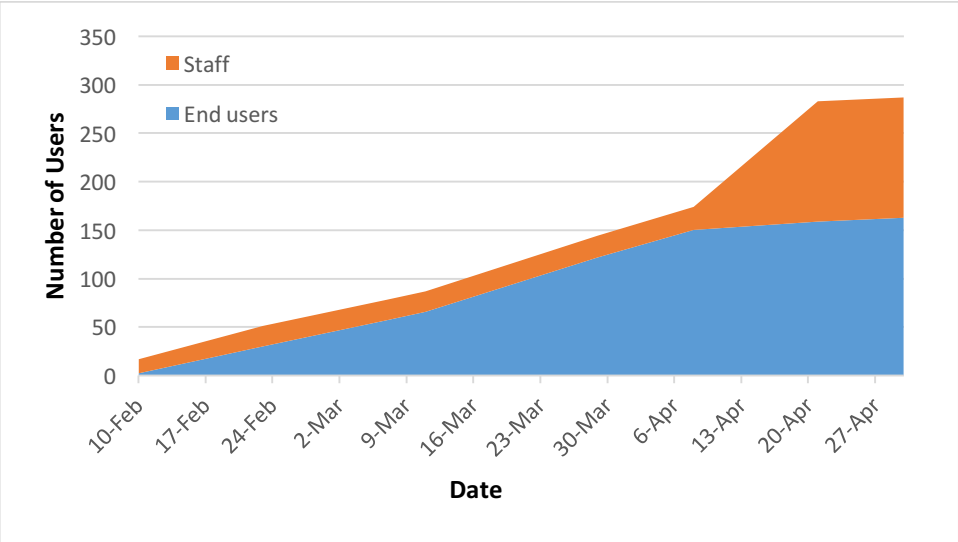


Figure 14. Number of distinct users who have used Jetstream since early operations began on February 10, 2016. This is a "stacked area" graph showing the total number of users in blue, and then staff users above in orange, so that one can easily see the total number of end users and the total number of users. As of April 30, 2016 there are 163 end users, and 114 Jetstream and XSEDE users, for a grand total of 287 individuals who have used the Jetstream system. There are an additional 758 Galaxy users of Jetstream who are not included in the count.

5.6. User survey and testimonials

Initial feedback from early allocation holders is positive. A brief summation of early feedback from allocation holders is presented in Appendix VIII.

Dr. Carl Boettinger (Assistant Professor in the Department of Environmental Science, Policy, and Management at UC Berkeley, http://www.carlboettiger.info) was mentioned earlier in this report as regards a tweet about Jetstream. He has been a user of the NSF-funded Chameleon experimental cloud system for some time. He started using Jetstream during the Early Operations period, and offered the following testimonial:

> *The cloud-style virtualization in both Jetstream and Chameleon have dramatically expanded the fraction of my projects that are benefiting from HPC resources. Instead of spending days or weeks setting up the environment and adapting code to run on our campus cluster I can spin up a virtual instance on your platforms any time I need to scale something beyond my laptop and have it running in minutes, since I can bring my own docker container. Likewise I can have a post-doc or undergrad running on XSEDE via a Jupyter or RStudio-server*

web interface rather than spend weeks teaching them shell & ssh before they can start working. That is really very satisfying. While something similar is possible on commercial clouds, their payment model makes that feel more risky for very experimental work, where it is difficult to estimate runtimes and associated costs. As a new faculty member building a lab this has been particularly important to me.

As a DOE CSGF grad student fellow I was introduced to classical HPC but often wondered what I was failing to learn to take better advantage of those systems. Today I believe that you are now supporting the 'long tail' of science and I think there's a wealth of new research to be done in enabling research in the big space between the laptop and classical HPC.

6. Demonstrated practical value of Jetstream demonstrated by results already derived by the US science research community using Jetstream

Information presented in the prior section is indicative of the potential utility of Jetstream as a first-of-a-kind CI resource provided by the NSF for use by the national research community and as a practical resource supporting meaningful scientific research. In other words, not only could Jetstream be used to support the national research community, the national community has already successfully used Jetstream to achieve new scientific results.

While some of the results reported here are incremental, all will be used in or in support of a technical publication submitted to a peer-reviewed journal or conference according to the scientists who have contributed them.

Below we outline several exemplars of early science results derived from use of Jetstream: fish evolution and biodiversity, snake evolution, plant evolution, computer and computational science (two). One of these is part of doctoral dissertation research by a graduate student. We note in particular that genomics and field research – two of the areas we wrote about as focus areas for use of Jetstream – are highlighted in this section. These useful incremental results help show the practical value of Jetstream to these communities – communities which are not traditionaly major users of other XD program resources. In addition, we describe one use of Jetstream in a university class (credit bearing) in computer science.

6.1. Biological science research

6.1.1. *Biodiversity conservation (genomics and field biology) and evolution (fish and snake) - Marlis R Douglas + Michael E Douglas, University of Arkansas (Fayetteville)*

Research in the Douglas Lab is best summarized as biodiversity conservation (see Figure 15). Patterns of biological diversity are identified using population genomic and phylogenomic approaches and then examined across landscapes and within the context of evolution to identify the ecological processes that drive diversification of natural communities. Another focus is on impacts of global change on small and isolated populations, particularly in the arid Southwest of

North America. This region is recognized as a biodiversity hotspot, with most species endemic to the region (i.e., only found in this area and nowhere else on earth), and many threatened or endangered. Next-generation sequence (NGS) technologies provide unprecedented opportunities to generate genome-scale data for natural populations of non-model organisms to assess biodiversity. Custom bioinformatics scripts and high-performance computing resources (clusters) are needed to processes these large data sets and statistically analyze them using Maximum Likelihood and Bayesian algorithms, which are computationally intensive (e.g., MCMC with millions of iterations for convergence).

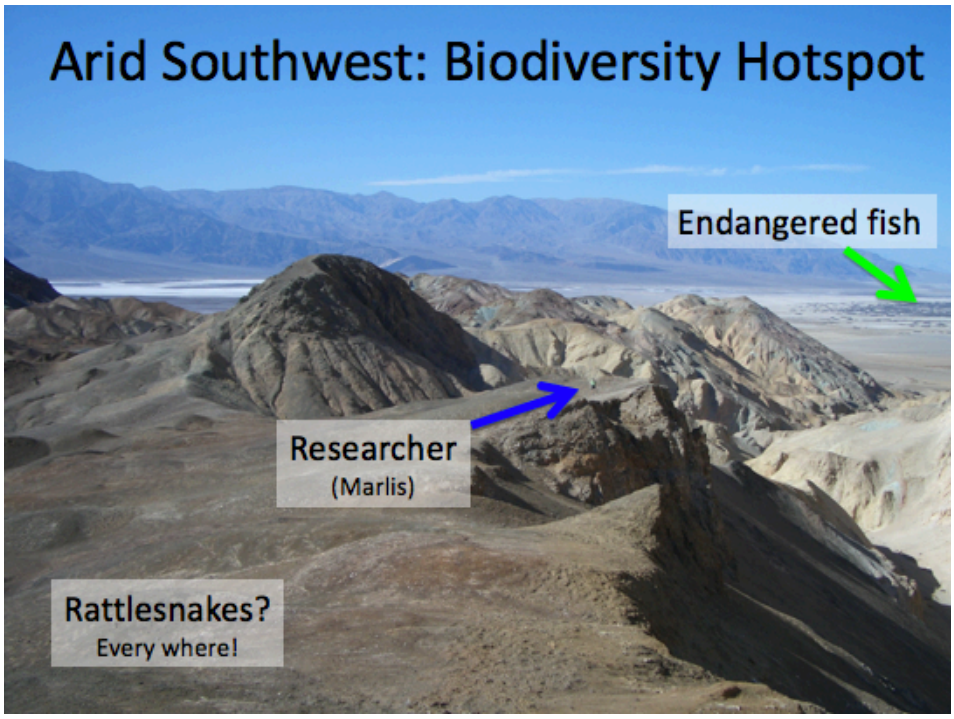


Figure 15. The setting for the fieldwork of the Douglas lab, in the arid Southwest.

Two ongoing research projects are currently using the computational environment provided by Jetstream: phylogenomics of Western Fishes; and Phylogeography in the Western Rattlesnake complex (*Crotalus viridis*). In both cases the genomic methods are as follows:

- Genomic methods
    - Extract genomic DNA from snake blood or fish fins
    - Cut genome in small fragments (~ 500 base pair)
    - Subsample ~40-60K fragments to reduce genome size
    - Sequence fragments with Next-Generation-Sequencing
    - 1 Illumina HiSeq lane: generates ~140 million reads
- ddRAD (Double Digest Restriction Associated DNA)
    - 6-8 HiSeq lanes / project
    - 400-800 samples of fish or snakes

In sum: LOTS of data and lots of analyses needed

6.1.1.1. Fish evolution

The Humpback Chub (see Figure 16, please note that this particular fish was released alive and unharmed) is one of the species endemic, endangered fishes studied by the Douglas Lab. Figure 17 shows an unpublished phylogenetic tree of suckers of the western United States. The portion of the tree highlighted includes the species being studied by the Douglas Lab.

Figure 16. One of the western fishes (released alive and unharmed) analyzed with Jetstream.

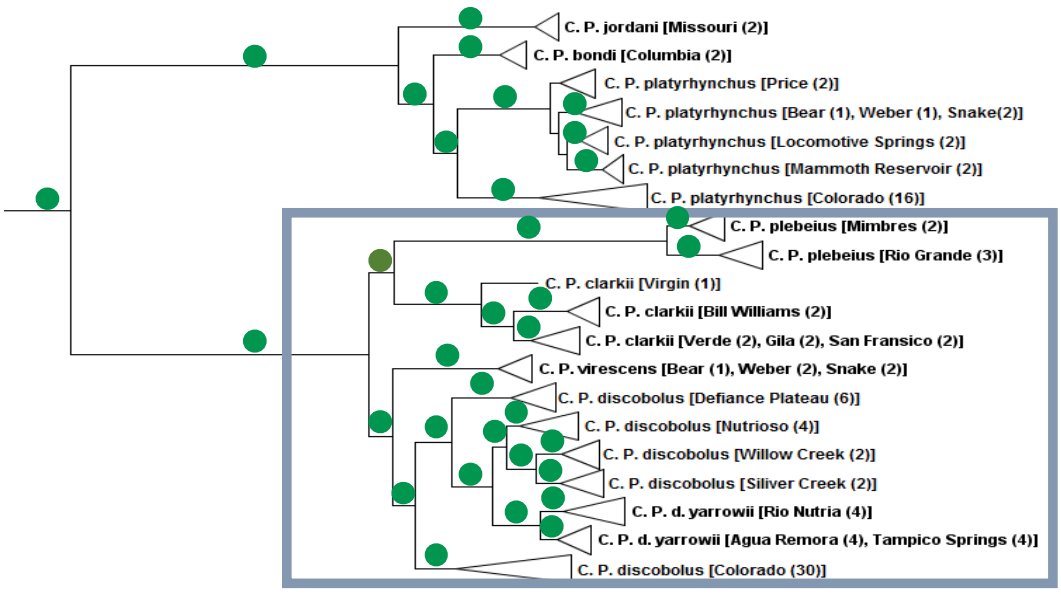


Figure 17. Phylogenetics of western suckers based on ddRAD data (unpublished)

The Douglas Lab analyzed ddRAD data via Bayes Factor Delimitation (BFD). This allows them to compare phylogenetic hypotheses incorporating thousands of loci integrating many gene tree topologies. This is computationally intensive work involving millions of iterations. Table 15 shows the results, which suggests distinct species among groups of suckers sometimes thought to be different populations of a single species.

Table 15. Bayes factor delimitation results calculated with Jetstream.

Model	Species	Marginal L	BF	10.79	Rank	Mean ESS
virgin+lcr+colorado	2	-3967.26	-	-	5	748.62
virgin, lcr+colorado	3	-3746.59	220.67	10.79	2	707.51
lcr, virgin+colorado	3	-3819.61	147.65	9.98	4	696.74
colorado, virgin+lcr	3	-3778.83	188.43	10.47	3	697.3
colorado, virgin, lcr	**4**	**-3642.89**	**324.37**	**11.56**	**1**	**702.22**

Other fish evolution analyses now underway with Jetstream include Bayesian species delimitation for Bluehead sucker. These models are significantly more complex with runtimes >20 days per model with multithreading.

Overall the work so far indicates reticulate evolution (evolution through the partial merging of two ancestor lineages, leading relationships better described by a network than the traditional bifurcating tree). This is in contrast to the neat directed acycling graphs. The Douglas Lab plans to submit their research, aided by analyses done with Jetstream, to the journal *Molecular Phylogenetics and Evolution.*

6.1.1.2. Snake evolution

The Douglas lab is also studying snake evolution (they specialize in scaly things, slimy or not). One current research goal in studying snake evolution is to clarify biodiversity within the Western Rattlesnake and, in particular, determine the number of distinct species in the *Crotalus viridis* complex by comparing genome-wide variation in SNPs (single nucleotide polymorphism).

The Douglas lab analyses ddRAD studies of DNA samples from 48 individuals representing 9 'subspecies.' At roughly ~ 25,000 loci / sample this came to ~ 4 GB of sequence data. The analyses done on Jetstream included:

- Genomic fragments aligned + clustered (pyRAD)
- Tested 3 clustering thresholds (0.85%; 0.90%; 0.95%) Lower values = more missing data allowed

- Phylogenetic Tree constructed (RAxML[4]) to compare topologies (patterns) and support values (statistical confidence) across different trees

The preliminary results are presented below in Table 16.

Table 16. Preliminary results of evolutionary relationships within the *Crotalus viridis* (Western Rattlesnake) complex

Clustering Threshold	# Loci	# SNPs
0.85%	24,176	127,850
0.90%	25,434	126,606
0.95%	25,548	95,326
0.85%	24,176	127,850
0.90%	25,434	126,606

The initial results suggest greater distinctiveness of subspecies within the complex than previously through (cf. their phylogeny based on morphology (e.g. [30]). The Douglas lab plans to conduct more analyses on Jetstream to improve the resolution of these analyses involving a larger data set (more snakes), testing other alignment/clustering parameters, and identifying potential hybrids. They plan to publish this research in the journal *Molecular Ecology*

6.1.1.3. Broader impacts of Douglas Lab research

In the proposal to the NSF to create Jetstream, the proposing team specifically targeted field biology and, in particular, biological research related to human impacts on the global climate and local impact on biological systems. The Douglas lab work is a perfect example of this sort of important research with broader impacts related to our ability as humans to co-exist with a healthy and sustainable global environment.

The Douglas Lab work on fish is focused on endemic species with adaptations to desert rivers, populations of which have declined due to habitat alterations (dams, stocking of predators, water diversions). Most of the species they study are rare, threatened or endangered, and the results of their studies help inform water conservation and water use decisions in the western US.

The Douglas Lab studies of Western Rattlesnakes are important in informing conservation efforts. This research also promotes an important societal good: surviving a rattlesnake bite if you are bitten. Surviving a bite depends upon correct species identification, development of species-specific anti-venom, and then stocking of appropriate anti-venom in reference to local snake populations, so the correct anti-venom is available to treat bite victims.

[4] Stewart notes with some pride that RAxML is descended from the parallel version of fastDNAml that his group distributed around the turn of the century - Stewart, C.A., D. Hart, D. K. Berry, G. J. Olsen, E. Wernert, W. Fischer. 2001. Parallel implementation and performance of fastDNAml - a program for maximum likelihood phylogenetic inference. Proceedings of SC2001, Denver, CO, November 2001. http://portal.acm.org/citation.cfm?id=582054

6.1.2. *Biology (plant evolution) - High Throughput De Novo Genome Assembly and Analysis. Dr. Eric B. Knox, Department of Biology, Indiana University*

The endosymbiotic origin of plastids and mitochondria from formerly free-living prokaryotes created three genomic compartments in plants, which have different evolutionary properties. Next generation sequencing of total DNA samples provides low-cost recovery of the genome components with many copies per cell, which are the complete plastid genome, the complete mitochondrial genome, and nuclear ribosomal RNA gene cluster. Jetstream is being used to assemble the millions of next gen DNA sequences per sample into the finished genomes, and to compare these results against a growing library of finished genomes for purposes of quality control and to detect evolutionary novelties.

An example of the family Campanulaceae from Africa

During the past year, the Knox Lab has sequenced the plastid genomes from more than 150 species in the plant family Campanulaceae, one-third of which are being sequenced more deeply to obtain the less abundant mitochondrial genomes and the nrRNA alleles. Knox uses use the genome sequences to determine the phylogenetic relationships among the species, and we use the phylogeny to analyze interesting aspects of molecular evolution. The Campanulaceae are unique among plants because the plastid genome has been repeatedly invaded by 'foreign' genes that probably originate in the nucleus, and in two sub-groups the mitochondrial genome is vastly enlarged, fragmented into many separate circular chromosomes, and has a fluctuating mutation rate. When all three genomes yield the same phylogenetic relationships, speciation was a simple process of lineage splitting (= cladogenesis), but we are also finding recent and ancient instances of reticulate evolution (= species of hybrid origin). With refined techniques, we have recovered genomes from small quantities of old tissue samples taken from herbarium specimens of species that are now extinct or very rare and difficult to obtain. During the coming year we will process an additional 200 species and delve more deeply into subpopulations of species with the most interesting results.

Plant mitochondrial genomes are typically 4-500 kbp long and organized in a single master circle. The mitochondrial genome of the African endemic plant Dielsantha galeopsoides (Campanulaceae) is 1,123,567 bp and is organized as five separate circular chromosomes. The increased size and fragmentation of the Dielsantha galeopsoides mitochondrial genome into separate circular chromosomes is consistent with other related species in this clade. The mitochondrial genome in these related species is as large as 2.7 Mb and fragmented into as many as 27 separate circular chromosomes, so Dielsantha galeopsoides (with only five separate

circular chromosomes) demonstrates that the enlargement and fragmentation is occurring independently in different lineages. Comparative analysis of these mitochondrial genomes will indicate how this fragmentation is occurring, and will ultimately address how mitochondria are able to maintain equal copy numbers of these fragmented genomes.

Results of these analyses will be published in American Journal of Botany, Molecular Biology and Evolution, PNAS, and (hopefully) Science or Nature.

6.1.3. *Psychological & Brain Sciences – Dr. Franco Pestilli, Indiana University (Bloomington, IN)*

The Pestilli Laboratory is geared toward improving understanding of the function and structure of the human brain to inform us about mechanisms of brain plasticity that occurs during development and aging. Dr. Pestilli is currently focusing on two important methodological advances for human brain science: (1) Advancing methods for mapping individual brains. Modern neuroimaging methods are bringing investigators for the first time to be able to obtain in vivo measurements of the human brain with the precision necessary to track fine changes in brain structure and function within an individual over their life span. This research effort aligns within the current National interest in Precision Medicine. (2) Implementing effective systems for open neuroscience. Replicability in Psychological and Brain sciences has recently come to the attention of the scientific community. Interest in scientific replicability has promoted a cultural change where research is changing from a 'cottage industry' model to an open science model. The Pestilli Lab is using Jetstream to implement a new model for sharing scientific results. Current best practice for open science promote sharing code, data and papers. The IU team is implementing a new approach to open science. Using Jetstream the goal is to establish practice for scientific replicability and sharing. The IU team will advance these practices by establishing new methods to share code, data and associated computational environments using Jetstream.

So far, three papers are in preparation, and one DOI with code, scripts, and data has already been published (published DOI: dx.doi.org/10.5967/P9WC7G.). Papers in preparation are to be submitted to Nature: Scientific Data and Frontiers in Neuroinformatics.

6.2. Computer and computational research and education

6.2.1. *Computer Science - Building a cloud resource orchestration system with Mesos on Jetstream. Renan DelValle, Ph.D. Candidate, Madhu Govindaraju, SUNY Binghamton*

From the paper by Christina Delimitrou and Christos Kozyrakis entitled "Quasar: resource-efficient and QoS-aware cluster management" (http://dl.acm.org/citation.cfm?id=2541941) http://web.stanford.edu/~cdel/2014.asplos.quasar.pdf we extract the following problem statement:

> *Cloud operators can achieve economies of scale by building large-scale datacenters (DCs) and by sharing their resources between multiple users and workloads. Nevertheless, most cloud facilities operate at very low utilization which greatly adheres cost effectiveness [9, 51].*

Utilization estimates are even lower for cloud facilities that do not co-locate workloads the way Google and Twitter do with Borg and Mesos respectively. Various analyses estimate industry-wide utilization between 6% [15] and 12% [24, 59]. A recent study estimated server utilization on Amazon EC2 in the 3% to 17% range [38].

Selected References

[9] L. Barroso. Warehouse-scale computing: Entering the teenage decade. ISCA Keynote, SJ, June 2011.

[15] McKinsey & Company. Revolutionizing data center efficiency. In Uptime Institute Symposium, 2008.

[24] Gartner says efficient data center design can lead to 300 percent capacity growth in 60 percent less space. http: //www.gartner.com/newsroom/id/1472714.

[38] Host server cpu utilization in amazon ec2 cloud. http://huanliu.wordpress.com/2012/02/17/ host-server-cpu-utilization-in-amazon-\ ec2-cloud/

[51] Charles Reiss, Alexey Tumanov, Gregory Ganger, Randy Katz, and Michael Kozych. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In Proc. of the Third ACM Symposium on Cloud Computing (SOCC). San Jose, CA, 2012.

[59] Arunchandar Vasan, Anand Sivasubramaniam, Vikrant Shimpi, T. Sivabalan, and Rajesh Subbiah. Worth their watts? an empirical study of datacenter servers. In Proc. of the 16th International l Symposium on High Performance Computer Architecture (HPCA). Bangalore, India, 2010.

The goal of graduate Student Renan DelValle's dissertation research is to build orchestration systems that utilize the ease of cloud systems and increase the efficiency of their use. Mr. DelValle has created an orchestration tool that makes use of the Mesos cluster management tool to easily create and use virtual clusters within cloud systems. Mr. DelValle demonstrated this system on the show floor, in the IUPTI display, at SC15.

6.2.2. *Project Aristotle – VM interoperability among multiple clouds*

There are now several different commercial and higher education public cloud systems in existence. Two public sector public clouds are Red Cloud at Cornell University and Jetstream.

Project Aristotle is a new NSF-funded project to create interoperability among different cloud systems. David Lifka, PI of project Aristotle (and a collaborator on this project) has stated: "The goal of the Aristotle Cloud Federation is to develop a federated cloud model that encourages and rewards institutions for sharing large-scale data analysis resources that can be expanded

internally with common, incremental building blocks and externally through meaningful collaborations with other institutions, commercial clouds, and NSF cloud resources." According to the press release about the project [31]. “The project name—Aristotle—was chosen because Aristotle’s concept ‘the whole is greater than the sum of its parts’ reflects the multi-institutional synergy and collaborations that the federation aspires to create.

One of the critical and practical goals of project Aristotle is to create mechanisms by which an application can be developed and used to the extent practicable on resources such as Cornell’s Red Cloud or Jetstream, and then moved to a commercial cloud such as Amazon Web Services.

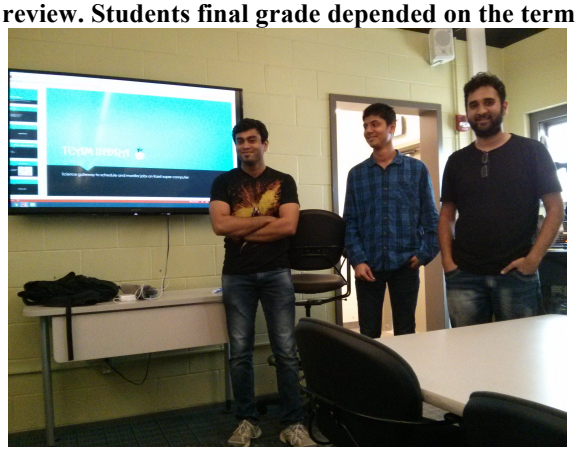
Dr. Robert VanRenesse has a startup allocation on Jetstream and has demonstrated with early tools developed by Project Aristotole the ability to migrate a VM among Amazon Web Services, Red Cloud, and Jetstream. That is, a VM has been instantiated on one of these resources, quiesced and moved to a second, quiesced and moved to the third, and then back round to the starting point… and working properly at each step along the way.

6.2.3. *Computer Science educational use. Marlon Pierce and Suresh Marru, Indiana University*

Students of the School of Informatics and Computing’s I590 graduate class, “Science Gateway Architectures” are using Jetstream to learn about cloud computing, microservices architectures, and “DevOps” software engineering principles such continuous integration and continuous delivery. The class is instructed by Research Technologies' Marlon Pierce and Suresh Marru; thirteen Computer Science graduate students are enrolled. The students are divided into four teams, each of which is developing from scratch a complete science gateway system. All code for each team is open source and managed in GitHub; students learn open source community practices as well as cloud-based approaches to building systems. Students’ final projects will feature complete science gateway software systems, developed as multiple interacting services, that get automatically deployed onto Jetstream VMs.

Students have been able to do meaningful and productive development of science gateways using Jetstream. The students of this class gave their final presentations in live demonstrations on May 4, 2016 (see Figure 18).

Figure 18. Students using Jetstream in a team project demonstration at the same time as the acceptance review. Students final grade depended on the term projects.



7. Production readiness: Operational-quality operations and early operations experiences as compared to management and operations metrics for Jetstream

In prior sections, we have asserted that Jetstream as implemented is indeed the system we proposed to build, that there are strong indicators of its utility to the scientific community, and we have demonstrated that it has indeed been used already to provide new scientific results.

In addition to being a first-of-a-kind system, we intend it to be a production system - a system that researchers and research students can count on being available when they want to use it. There are two components to this consideration relevant to consideration of acceptance of Jetstream by the NSF: production operations and our early experiences operating Jetstream as indicators of successful fulfillment of Management and Operations phase metrics.

7.1. Production operations

IU and TACC have excellent physical infrastructure for housing advanced computing systems. We will not repeat here information that was contained in our proposal regarding the IU and TACC facilities for housing advanced cyberinfrastructure. We will report here new security features put in place since the proposal was submitted, and comment on the actual use in practice of some of the management tools described in our proposal

Security is particularly crucial in a cloud environment where end users have escalated privileges and bring their own tools and software. IU and TACC have a long history of managing resources that are constantly under attack from malicious persons and groups. This experience has led to increased attention and cooperation with local security experts in architecting and deploying Jetstream, and with any system a layered approach is required. Featured VMs are, and will be, regularly updated and patched by skilled administrators to provide secure by default environments. Isolation of VMs through encapsulation with VXLAN and use of private IP addresses where possible provides another layer of security. Host based and hypervisor based firewalls and configuration restrictions are also used to further implement sound security practices. Traffic to Jetstream flows through load balancers at each site and provides a key point where Jetstream can tie into network intrusion detection systems (NIDS) that can implement automated blocking for malicious activity. Indiana University's NIDS consists of two clusters each with 18 worker nodes, a manager node, and a dedicated packet capture host for troubleshooting and in-depth monitoring of connections of interest. These clusters run the Bro network security monitor software and also use SNORT intrusion detection rules [32], [33]. The NIDS receives data from span ports from our Science DMZ (Jetstream-IU sits within the IU Science DMZ network), as well as taps on border routers, data center spine switches, and wireless controllers.

When compromises do happen they can also be detected by NIDS or network monitoring allowing Jetstream administrators to capture and stop the activity in an expeditious manner. Results of these issues will be reported through already established channels for Jetstream partners and the XSEDE security working group. Jetstream also leverages SSL protocols for OpenStack endpoints along with multi-factor authentication for privilege escalation on

administrative systems. As a result of the particular needs to ensure good security while operating cloud environments IU has recently invested in additional software monitoring tools to detect any signs of attack or access patterns that indicate suspicious program or user behavior.

We have, as we indicated in our proposal, made extensive use of software tools to facilitate collaboration and management of this project. In order of most used to least used, the critical collaborative and management tools used to support implementation of Jetstream are:

- Slack, used particularly to support real-time chat among the distributed team of programmers and systems administrators at IUPTI, TACC, and the University of Arizona. Slack is used many times an hour during business hours, and often outside of business hours.
- JIRA for problem and bug tracking of software implementations within the team. JIRA is used daily.
- GitHub as a software repository for both Atmosphere and Jetstream-specific deployment software.
- Request tracker for tracking user issues through the XSEDE trouble ticket tracking system. XSEDE's instance of RT is used daily for handling user issues and communicating with users.
- Confluence wiki for general group communication. Confluence is the one communication tool that all team members have access to. Most team members use it daily.
- NCSA's "Savannah[5]" risk management software. IU has a local installation of the "Savannah" risk management software developed by NCSA, used at IU thanks to NCSA and their willingness to distributed this software as open source. This tool is used primarily by the PI (who is the primary updater of risk management information so far). The PI uses this tool monthly (see Figure 19 for a screenshot).
- Projectmanager.com is used for high level WBS management by project manager Therese Miller. Ms. Miller uses it periodically; so far much of the work has been so fine-grained within the tasks of implementing the system, and the number of active funded partners small enough, that we have had little need to use this tool particularly actively.

[5] With regards to the Savannah risk registry system. There is one quirk in the software: the "Owner" field shows up as a blank for risks owned by the person logged in to the system. In this case every risk shown with a blank in the order field is actually owned by Stewart).

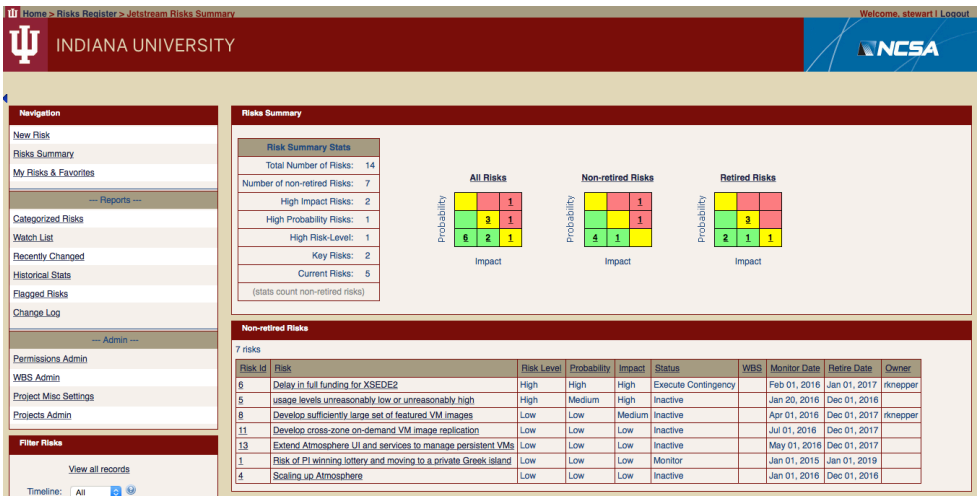


Figure 19. Screenshot of the NCSA Savannah risk registry system in use by PI Stewart as of April 23, 2016.

7.2. Early operations experiences relative to metrics defined for the Management and Operations phase of Jetstream

The Jetstream PEP contains metrics goals for Jetstream for its post-acceptance Management and Operations phase. These metrics are according to the PEP not part of the acceptance criteria for Jetstream. Still, it is of some use to consider metrics for the month of April in order to develop a better understanding of the usability and use of Jetstream. This helps inform understanding of Jetstream as a production system simultaneously with learning from this early experience with Jetstream in the sense of it being a pilot in the production cloud space for the NSF. Management and Operation phase metric targets are presented in Table 17.

The data presented below cover April 1 – 30, 2016. In this table the "Green" icon is used to indicate that the usability and use experienced during the month of April so far would, if continued out over the year, result in successful accomplishment of metrics for Management and Operations Program Year 1 as defined in the PEP. The Yellow icon appears in a few places and indicates areas where we are yet learning what the right metric targets are and evolving use of the system. The "White" (not enough data) icon appears in one place – number of publications. We feel very good about the number of analyses that have been done with Jetstream already, but feel it is premature to make predictions of success until there are at least a handful of manuscripts in review.

Table 17. Milestones for ongoing operations for the month of April 2016. These milestones are indicators of performance and use.

	Goal per program year	Pro-rated per month if appropriate	Achieved in month of April	Notes	Outcomes
Availability					
System availability (uptime of an element of the production hardware, as % of wall clock time)	95%		100%	95% (average annual target) is the required NSF measure of success.	
Capacity availability (% of the total capacity of Jetstream available for NSF use over time)	95%		100%	"	
Job completion success - % of jobs submitted should complete without having to be resubmitted as a result of a failure in the hardware or system software.	96%		97.7% of featured VM launches that were allowed by their user to complete reported to Atmosphere as active.	96% (average annual target) is the required NSF measure of success.	
Utility					
Core cloud environment software will be upgraded to match current versions components such as operations systems and cloud software environments	Updated prudently, generally within ≤ 12 months of major releases		We are running on the most recent version of OpenStack and Atmosphere available		
Utilization					
Capacity of system allocated via XSEDE	90%		NA	> 90% allocated, but this is not meaningful at this point since time used on Jetstream is not charged against user allocations during the early operations phase.	
Total number of distinct users	1,000	To achieve this in a year, need to add 84 users/month	327	These numbers represent aggregate running total targets	

	Goal per program year	Pro-rated per month if appropriate	Achieved in month of April	Notes	Outcomes
Total number of students having used Jetstream in an educational or training setting	100	Add 9 users per month	12	"	
Total number of science gateways using Jetstream	2	Add 1 gateway every 6 months	3	"	
Use - average number of VMs active 24 hour average	320		Average of 290; peak of 1217 (15-30 April)		
CPU % utilization	6%		Average of 4.2%; peak of 20.3% (15-30 April)		
Outcomes (Utility)					
Total number of publications facilitated by use of Jetstream	5	1 every two months	Multiple papers now in process	These numbers represent aggregate running total targets	
Total number of VM images and/or data sets published with a DOI via IU Scholarworks	10	Just under 1 per month	6	"	

7.3. Notes on early experiences relative to Management and Operations metric targets

Our early experiences suggests that we will be on track to meet almost all of the management and operations metrics targets set in the table above. So far the two areas that do not yet demonstrate a clear and sustained trend toward meeting Management and Operations phase metric targets are in the areas of "numbers of VMs in use" and CPU utilization. CPU and VM utilization are the two areas of Jetstream performance that we will continue to pay close attention to throughout our management and operations phases.

Figure 20 and Figure 21 show our empirically observed results on load at various points during the month of April.

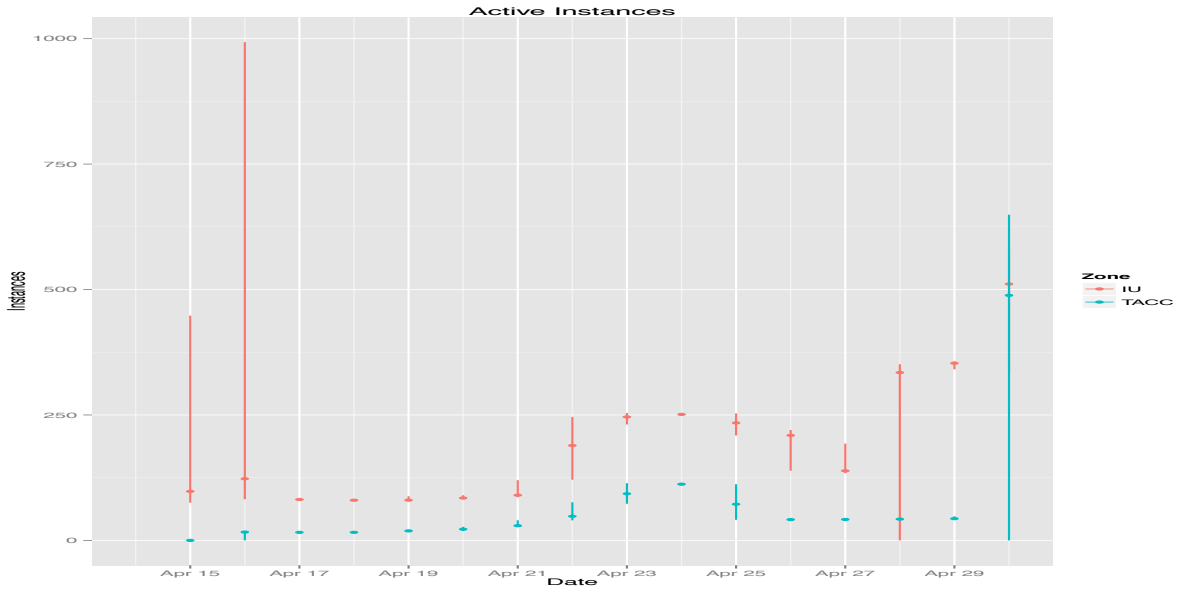


Figure 20. Min/Max/Mean totals of VMs in use on Jetstream-IU and Jetstream-TACC from April 14-30, 2016

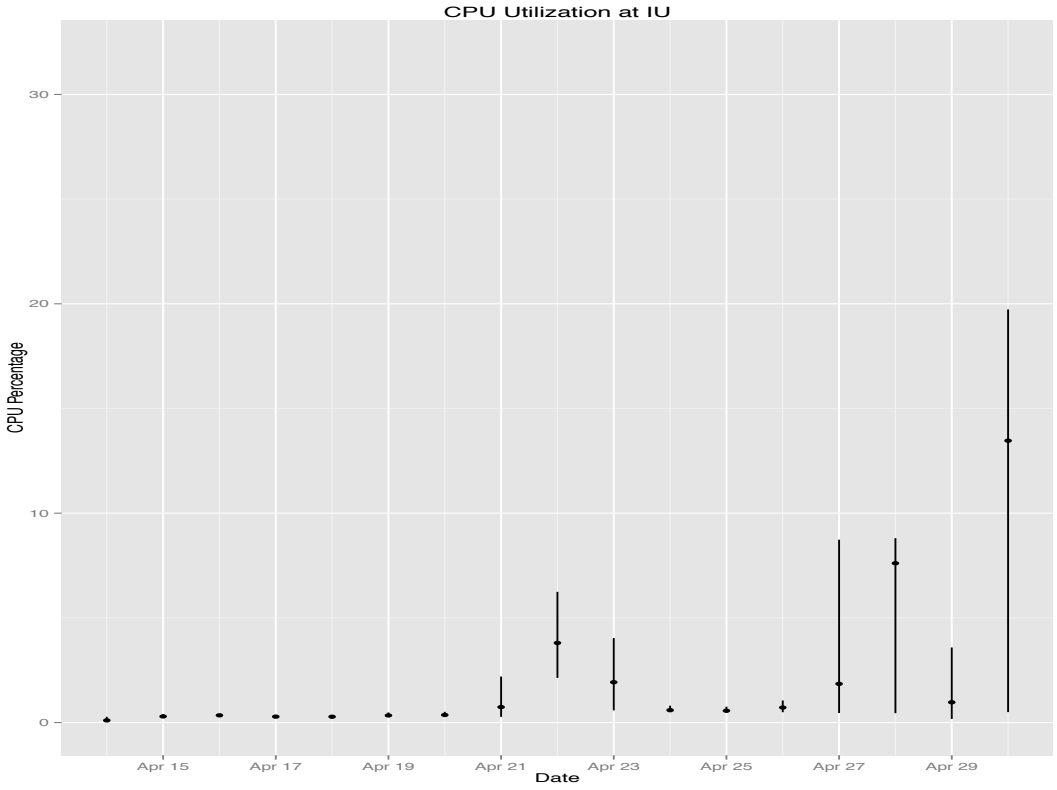


Figure 21. CPU load on Jetstream-IU from April 14-30, 2016. Data from April 1-13, 2016 were lost during a reconfiguration of the data collection settings.

Starting at noon EDT on the 21st of April, we had two different efforts of 24 hours each in which we asked staff to interactively launch as many VMs as possible by hand and by scripted execution of jobs from SEAGrid and Galaxy. Table 18. VM and CPU usage is shown for data we have in hand for the month as a whole, and for usage for the noon EDT April 21 to noon EDT April 23 as well as for April 30 is shown in Figure 22 and Figure 23.

Table 18. CPU load and VMs active on Jetstream

Metric	15-30 April	noon 21 April - noon 22 April	noon 22 April - noon 23 April	April 30
VMs				
Jetstream-IU & TACC combined	Average: 290.1; Max: 1217	Average: 208.9; Max: 315	Average: 330.0; Max: 364	Average: 1201.2; Max: 1209
CPU load				
Jetstream-IU	Average = 4.2%; Max: 20.3%	Average 3.03%; Max: 5.43%	Average: 2.96%; Max: 4.45%	Average: 13.38%; Max: 20.31%

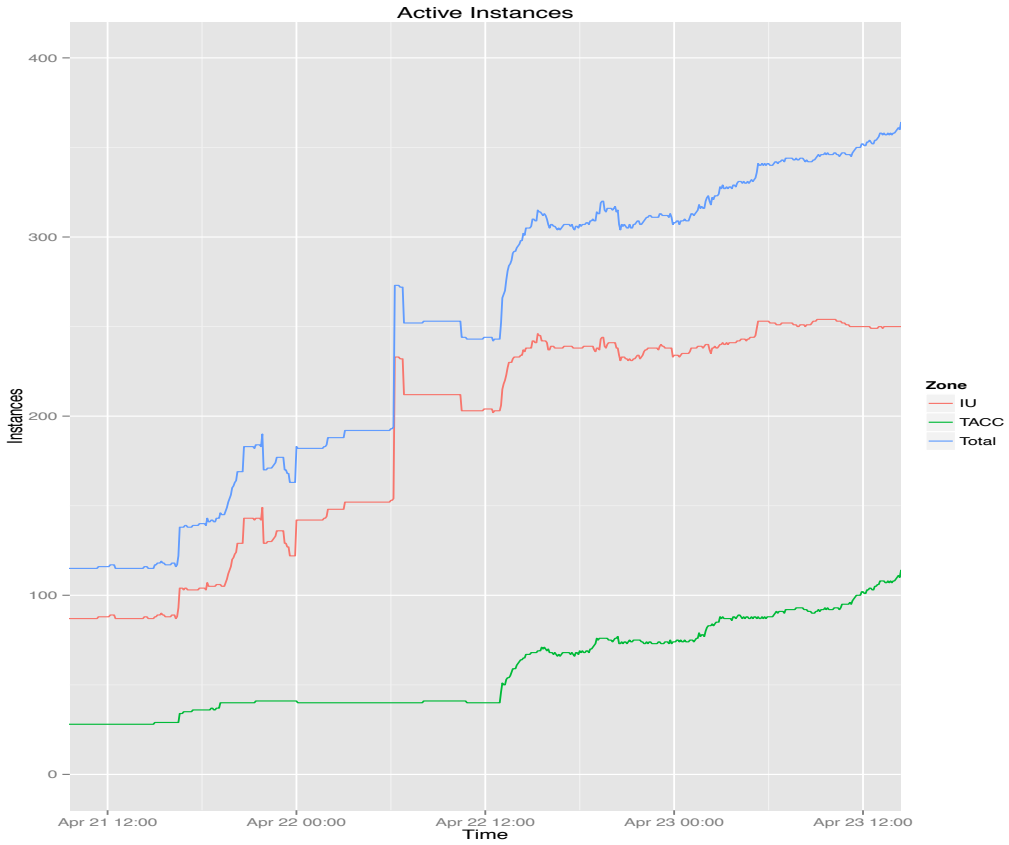


Figure 22. Number of VMs in use on Jetstream-IU and Jetstream-TACC for two days starting 12:00pm EDT April 21, 2016.

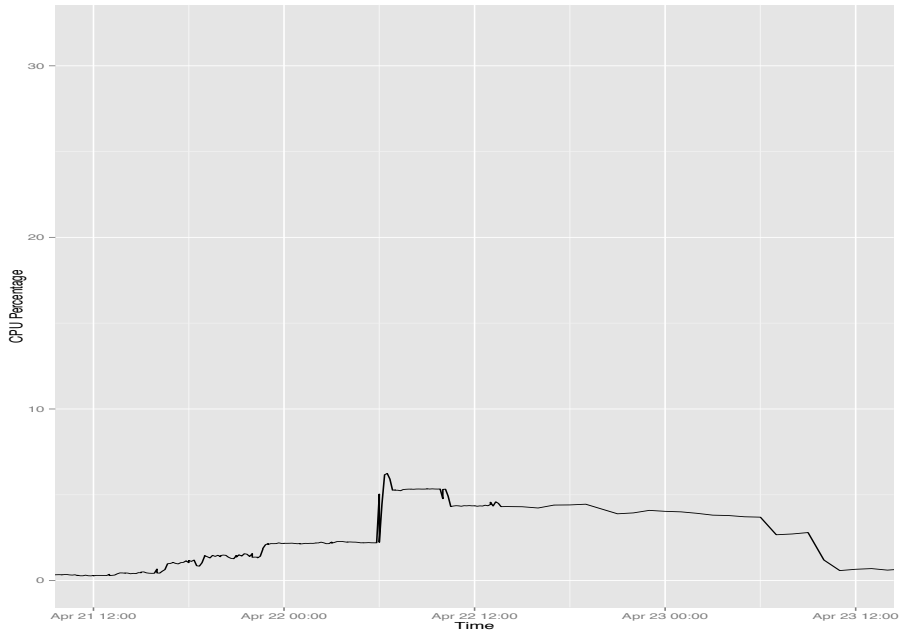


Figure 23. CPU load on Jetstream-IU for two days starting 12:00pm EDT April 21, 2016.

Our starting points in designing Jetstream and planning for workloads were:

- Focus on interactive users, which represents the vast majority of the need expressed in our reading of pre-existing survey results conducted by other researchers and in our own interviews with researchers as we developed use cases.
- Put a premium on responsiveness of the system to individual users so that the user experience seems positive and consistent. (Whether the observation and inference we heard is correct or not, the observation made to us by several users we interviewed was that their experience of commercial cloud environments was that performance was inconsistent, and their inference was that inconsistent patterns of oversubscription on the part of the commercial cloud operators).

During early operations we have had some very good experiences with a very excited group of users. Usage has grown organically during the early operations phase. During the first part of early operations we focused on increasing the number of users of the system. During the latter part of early operations, particularly the last three weeks of the month of April, we focused on aiding users of Jetstream in getting useful scientific research done and thus the rate of increase of users was slower during this time period.

What we have learned so far in our early operations experience with Jetstream is as follows:

- Loading VMs on to nodes at 1-2 VMs per node has resulted in good user experiences
- Interactive load at a ratio of 1-2 VMs per node results in a CPU utilization level that seems low when taken out of context. We understand that at present any statistic regarding any financial investment by any branch of the federal government will be taken out of context sooner or later.

- We can with synthetic loads demonstrated the capability to run many more than 640 VMs total on Jetstream - 998 VMs simultaneously on Jetstream- IU and 832 on Jetstream-TACC. A combined 1217 instances on both sites (with at least one VM on every node).
- We can generate a load that seems rational: more than 5% CPU load to more than 20%
- Support for the resource via standard XSEDE mechanisms. We have received 21 support tickets through the standard XSEDE trouble ticket system, and more than 80% of them were resolved in less than 1 week. This is not good enough for production services, but it is surprisingly close given that the system status is officially "early operations"
- We have also concluded that we are providing, for a resource intended to be an interactive computing system, CPU statistics that no other NSF-funded resources currently report publicly.

In the future, we plan to carefully monitor and manage outreach, training, and dissemination as well as monitor and adjust usage so that Jetstream will simultaneously be an effective interactive resource as originally intended and proposed, and also be a highly useful computational cloud system providing significant system and CPU resources to the nation (see Appendix IX). Our initial planned steps are as follows:

- Outreach, education, community building
    - Continue extensive outreach activities at domain science conferences (and minimize effort spent at conferences such as SC).
    - Fund, from the Jetstream Management and Operations budget, the creation of one Cornell Virtual Workshop about Jetstream per year, recognizing the value of such on-demand training and the limitations of XSEDE capacity to develop such training
    - Continue our focus of leveraged support working with Virtual Organizations and Communities of Practice
- Load management
    - Continue a focus on optimizing system management parameters for an excellent interactive user experience
    - As experience suggests that we may, increase oversubscription of VMs to nodes
- Load generation. Focus on API-driven access to Jetstream during non business hours to make excellent use of computational resources. Early exemplars of applications and user communities that will take advantage of Jetstream as a resource include the following:
    - SEAGrid – already now able to use Jetstream. We will promote use of Jetstream as a SEAGrid resource as soon as the system is formally accepted.
    - Galaxy – Galaxy main at usegalaxy.org is now able to route computational work to Jetstream. We will promote use of Jetstream as a SEAGrid resource as soon as the system is formally accepted.
    - ATLAS – we have an allocation in place to accept jobs on Jetstream to do computational work analyzing ATLAS experiment data. Some preliminary work has been done, and this will be one of our two top priorities in hardening production access to Jetstream once acceptance is complete.
    - iPlant/CyVerse – Nirav Merchant's group at the University of Arizona and Douglas Thain's group at the University of Notre Dame have created software tools that will dispatch iPlant workloads to remote cloud resources. This toolset

remains mostly unused in production for lack of a cost-free resource to use as a remote computational facility, and lack of funding sufficient to use Amazon Web Services. With the ATLAS work mentioned just above, once we are in production with Jetstream implementing this tool to support iPlant jobs programmatically through the Atmosphere interface will remain one of our top two initial priorities in supporting meritorious research and at the same time making good use of the CPU resources within Jetstream.

- Research and development with XSEDE and the State University of New York (SUNY) at Buffalo:
    - We have already established plans to work with SUNY Buffalo on the measurement software XDMoD (XD Metrics on Demand). We plan together to implement XDMoD on Jetstream in ways that measure CPU utilization from a "within the VM" viewpoint as well as at the overall system level. This will help us better understand the use of the system overall.
    - We have come to the conclusion that the cyberinfrastructure community talks too little about CPU utilization overall. Comments such as "The system is 90% busy" are made at the level of nodes being reserved for a job or queue slots that are busy. Even we in our early discussions allowed ourselves to be distracted by figures like these. We intend to pursue a dialog generally to better understand and explain how systems overall are utilized so that the NSF and the CI community can properly put in context the very different concepts of CPU utilization overall at the system level, CPU levels within a VM or within the execution of a particular job, and system business generally in terms of percent of nodes busy, or number of maximum percent of VMs in use.

8. Jetstream as a cyberinfrastructure resource

In this section we expand on the role of Jetstream as a cyberinfrastructure resource within the context of XSEDE and the NSF XD program. Here, we focus on what it means for Jetstream to be a cloud managed for science, and on Jetstream as a production service that services both a pathfinding (pilot) role for the NSF and which will be, if the NSF so chooses, scalable over time.

8.1. Jetstream is a managed science and engineering cloud – a cloud managed for science and engineering

Indiana University pioneered in the 1990s what it referred to as the "leveraged support model," in which a central Information Technology (IT) provisioning organization leverages a suite of self-serve online help resources and community partners to deliver more support to a large user community than the IT organization could ever deliver directly [22], [23]. Leveraging information via the Web now seems old hat; it was not when IU released its online Knowledge Base in 1996 as a production support tool.

Today, web-based tools and social media allow much more sophisticated interactions between support communities, end users, and an array of members of communities of practice (CoPs) and

Virtual Organizations (VOs). Our implementation and support strategy for Jetstream is based on collaboration with formal and informal groups of professional cyberinfrastructure experts, expert users, instructors, students, and new users as well as the staff of XSEDE (the eXtreme Science and Engineering Environment).

With regards to help from staff of XSEDE, we are pragmatic in our expectations. We expect expert and valuable help from the allocations staff and processes of XSEDE. In the early days of Jetstream operations this may be the most critical contribution of XSEDE staff to Jetstream operations. We expect significant assistance in training but sufficiently limited in scope that we have built the creation of Cornell Virtual Workshops into the Management and Operations budget for Jetstream. We expect, over time, excellent and important assistance from XSEDE Extended Collaborative Support Services in in-depth performance tuning and science gateway development. We expect little to no practical help from XSEDE in resolving day-in, day-out user problems other than authentication and accounting problems.

The day-in, day-out support of Jetstream will fall largely on the CoPs and VOs that are our partners in developing the concept of operations for Jetstream. Some of these partners will receive modest amounts of funding in the planned Jetstream Management and Operations phase of Jetstream. Most will not. Indeed, a significant amount of the support for Jetstream will be done by partner institutions, CoPs, and VOs because the combination of Jetstream's user-friendly interface and existing support structures within scientific communities provide a faster and better path to scientific results – for the user groups we target as users of Jetstream – than other existing cyberinfrastructure resources. Indeed, already during the friendly user phase and early operations phase we have seen the XSEDE Campus Champions and their email list serving as a vehicle for dissemination of help information about Jetstream. Once Jetstream is in operations we expect to make significant use of social media (interactive wikis, Facebook, Twitter) as ways to communicate with and among Jetstream team and its many (and sometimes overlapping) communities of users.

Another important point goes back to the vision of Jetstream as a managed science and engineering cloud – a cloud managed for science and engineering. Open science and engineering research needs and concerns for broader impacts drove the design of Jetstream. This drove our disciplinary focus, the system design, and the way we developed concepts for contributed VM images.

The Jetstream system and the team implementing Jetstream are built around attention to NSF priorities in broader impacts. This is evident in our focus on biology – including field biology – and topics relevant to understanding the impacts of global climate change. We have also specifically built our broader collaborations around partnerships that help build diversity into the Jetstream user base.

Table 19. Partners in deployment and support of Jetstream

Discipline or mode of use	Lead partners	Funded?
Biology	iPlant, University of Arizona, Galaxy, Johns Hopkins University, Penn State	Primarily for tool delivery; support of biological research

Discipline or mode of use	Lead partners	Funded?
	University; genome analysis	community per se is not funded
Earth Science/Polar Science	National Snow and Ice Data Center (NSIDC), "High-Performance Distributed Computing in the Polar Sciences" Research Coordination Network (RCN)	No
Field station research	University of Arizona (Bryan Heidorn)	No
Geographical Information Systems	IU	Not for this aspect of activities
Network Science	IU Network Institute	No
Observational astronomy	WIYN Consortium	No
Social Sciences	Odum Institute, University of North Carolina	No
Campus Bridging	XSEDE, Cornell, IU	Not from Jetstream budget
Outreach to schools with limited budgets – including HBCUs and schools in EPSCoR states	University of Hawaii, University of Arkansas Pine Bluff	No
Use of proprietary software	Mathworks	Not from Jetstream budget
Facilitate reproducible data analyses	University of Chicago Computation Institute	Not for this activity
Enhance Science Gateway Deployment	University of California San Diego (San Diego Supercomputing Center), XSEDE	Not from Jetstream Budget
Visualization and analysis	IU, University of Texas	Not from Jetstream budget

8.2. Jetstream is scalable and a valuable learning experience for the NSF and the national research community

For several years the open science research community has called on the NSF to deliver cloud resources for use by that community. The NSF has now funded three different cloud and grid resources for experimental computer science research (Jetstream, Chameleon, and CloudLab). Jetstream is a first-of-a-kind deployment as an NSF-funded cloud environment intended for production use by practicing scientists.

The deployment of Jetstream has been and continues to be a learning process for the deploying team and for the NSF. A great deal of work has been done understanding how to integrate the OpenStack cloud environment with NSF-funded software environments such as Globus and Atmosphere. A great deal of work has been done (and is being disseminated) regarding OpenStack deployments in a university setting (It's great, but not an experience quite like one reads about in the trade journals and blogosphere). A tremendous amount of learning has already been done by the Jetstream team with NSF experts trying to adapt metrics developed over decades (and in some cases longstanding requirements of HPC acquisition solicitations) to a cloud setting. The work of deploying Jetstream as a managed science cloud – managed openly

for science with lessons disseminated to the scientific community – is adding significantly to our collective understanding of cloud technology.

Every first-of-a-kind computer system is a bit of an experiment. "Can we make it run, and will people who have said they wanted it actually use it if we do?" is very definitely a social experiment. In some cases, we have already completed successful experiments (or at least, completed software engineering implementations) that will be ensconced in open code repositories and described in technical papers. The Jetstream team and the NSF have already come to new conclusions with yet open questions about accounting practices and metrics. The value of such an experiment – is a cloud funded by the NSF really going to be valuable to the national open research community – accrues best and most profoundly over a number of years. Over years we can begin to understand patterns of use, understand the importance of discoveries made with the Jetstream system that are not practicable with other existing NSF-funded CI resources.

It is a challenge to quantify the value of a cyberinfrastructure resource such as XSEDE, particularly when such resources often support basic research that may yield societal benefits over the course of years or decades. Stewart and his colleagues have been leaders in this area, publishing what seems so far to be the only analysis of Return on Investment (ROI) of a cyberinfrastructure resource that exists in the peer-reviewed literature [34]. Matt Link of IU has funding from the NSF, with a subcontract to SUNY Buffalo, to expand XDMoD to include features to automate some aspects of measurement of ROI. We will publish an analysis of ROI relative to NSF investment, comparing it to alternatives such as use of commercial clouds, at the midpoint of the four years of operations of Jetstream.

Several reports and many individuals have suggested that the NSF fund a cloud resource. Still, it's worth asking at this point once again: why not just depend upon the private sector? We believe the following factors – many of which are based on our early experience to date, suggest that it makes sense for the NSF to fund a resource such as Jetstream. It is, as we put it, a managed science and engineering research cloud; a cloud managed for science and engineering research as its first priority. Commercial clouds are not driven primarily by science needs. Why should the NSF invest in this?

1. It is without cost to the end user, via an allocation process. That encourages scientists and engineers – particularly scientists and engineers in domains that have not traditionally made deep use of advanced cyberinfrastructure – to use it.
2. We work from the user interaction layer of Atmosphere on down through OpenStack to network tuning, so we provide a degree of vertical integration, testing, and tuning that would be harder to manage with atmosphere on a commercial system. In a sense, the Jetstream team provides for scientists and engineers working in the long tail of science the same sort of service that commercial cloud providers give their big customers (where "big" means "big like Apple" or "big like Pixar.")
3. We test, tune, and certify a set of VMs that we stand behind and guarantee to work, and we work with communities of practice, VOs, and disciplinary groups to prioritize what we make available as a "featured VM" to best meet community needs.

4. We provide a clearinghouse of VMs contributed by scientists who want to make them available to the research community.
5. We offer up for free the service of storing VMs (and data files and scripts and programs and output) as a digital object in a persistent digital archive (IUScholarWorks, running on DSpace / HPSS) and we assign DOIs to them. The storage archive behind this service has been in continual operation for 17 years at this point.
6. Because Jetstream is free to end users, and because we do a lot of handholding with researchers, we provide a place where they can convert their analyses to a cloud environment, scale up their work, get help, etc. and then when they get beyond what we can provide for free, they can take their VMs and money from someplace (home institution, NSF, whatever) and move to a commercial cloud.
7. The Jetstream operational characteristics encourage users to keep their own data someplace other than on the Jetstream cloud environment for the long haul, rather than using the difficulty of getting your data out of the cloud as a way to promote customer lock-in (as some commercial vendors seem to do, in order to maintain long term income streams).
8. There is inherent value in diversity, in terms of community stability. It is a well-known result from community ecology that diversity creates ecosystem community. This works in ecosystems of plants and animals and in human-created ecosystems as well. One of the factors in the bank crash of '08 was essentially no diversity in risk management: every major bank depended on the same risk sharing pool (credit default swap) so when things went bad all of the big banks had one common problem. Jetstream makes the community of cloud providers more diverse. Within the OpenStack community it is one of the larger government supported clouds. Diversity is inherently better than monoculture over a long enough time period because diversity means there are more sources to pick from when cherry-picking innovations.

There is inherent value in Jetstream as a pilot for the NSF. There is a very important bit of budget safety in this for the NSF. The usage of this system is metered by the capacity of the system they purchase as a capital investment. This makes it possible for the NSF to invest in a way that they can plan for when they otherwise would find it harder to plan and manage. Furthermore, after four or at most five years, the system will go away. It's not an open-ended investment. As a pilot, it allows the NSF to make a significant but bounded investment (around $15M over 5 or so years) and learn from this pilot project as the NSF decides what the role of cloud computing is in its cyberinfrastructure plans for the future.

We note that many of the benefits listed above could be achieved by funding the Jetstream team to work as a service organization and allocating a fixed amount of funding for resources on a commercial cloud service. But, not all of them. Thus, over the next four or five years of production operations of Jetstream, we expect three primary sorts of benefits:

- Many important scientific discoveries made by communities of researchers that are new to the XD program and have not used XSEDE-supported resources before
- Significant broader impacts stemming from use of Jetstream, ranging from workforce development to societally important outcomes of the work of the users of Jetstream

- Significant knowledge for the NSF and for the open science community about what are, in practice, the costs and benefits of running a cloud system rather than buying services on a commercial cloud.

As we learn more about the use and value of Jetstream, the NSF also has the opportunity to expand the scale of the Jetstream deployment. It is a well-worn adage that the human resources needed to administer clusters scales more strongly with the number of clusters than the size of the clusters. This seems especially applicable in our experience to cloud resources built with OpenStack. Expansion of the hardware resources at IU and TACC may be done incrementally as needs and NSF funds allow. This is one key part of the "scalability" of Jetstream promised in the proposal title. Another part of the scalability of Jetstream is in the collected knowledge and assembled open source code base that will allow other institutions to stand up OpenStack clouds that may or may not be integrated at the authentication and accounts level with the existing Jetstream resources. Last, and definitely not least, through implementation of Jetstream the participating organizations are dramatically scaling up the number of CI professionals working in higher education and open research laboratories that have first hand experience implementing cloud resources in general and OpenStack cloud resources in particular.

9. Conclusion: Jetstream is now implemented in a way that successfully fulfills the definition of Jetstream in the Cooperative Agreement and PEP

In this document, we have demonstrated that:

- Jetstream meets the requirements set out in NSF solicitation 14-536 in terms of integration with XSEDE.
- Jetstream fulfills the test specified in a peer-reviewed Project Execution Plan in ways that demonstrate that Jetstream as currently implemented is indeed the system proposed, and it operates successfully.
- Jetstream is being allocated as called for in NSF solicitation 14-536 as a resource at 90% of its theoretical capacity via NSF-specified allocation processes operated by XSEDE (the Extreme Science and Engineering Discovery Environment).

The above should be sufficient to declare the system accepted and ready to be put into operations.

As a resource intended as a production resource, it is worthwhile to also consider our experiences from friendly user and early operations modes. (The Jetstream system is one of the first systems in the history of NSF HPC acquisitions to be referred to as a production system by its proposers). From our friendly user and early operations mode we have learned the following:

- The system has utility to the national science and engineering research and research education community – in the sense of having a variety of software tools available that make it useful already to the majority of the user communities we identified as intended users of Jetstream.
- Jetstream has been used to support data analysis resulting in new scientific knowledge.

- Jetstream has been used directly by a total of 287 people, of them 163 “end-user researchers or students” (people who are employed neither by any part of the Jetstream team or XSEDE). An addition, 758 people have used Jetstream indirectly by submitting jobs to Galaxy through usegalaxy.org running on Jetstream.
- Jetstream has been used to perform meaningful scientific research. In this document we have included several short summaries of incremental scientific results that have already resulted from use of Jetstream.

10. Appendix I. Detailed timeline

A detailed system description follows, after which are details on the performance targets, the methods used to perform the acceptance tests, and the achieved performance. This timeline focuses on the production components of Jetstream (IU and TACC) rather than the entire project or the Jetstream-AZ (J-AZ) system, which has already been accepted.

The majority of the Jetstream-IU (J-IU) Production cluster was delivered to the Indiana University – Bloomington's Data Center on 10/19/2015. One compute rack was held at the Dell Merge Center due to a single top of rack S6000 switch that failed to pass Dell's acceptance criteria as well as four blades with non-functioning Ethernet network interface controllers (NICs). Additional issues were encountered plumbing the water cooling doors resulting in the inability to power on the J-IU system in its entirety until 01/14/2016. Otherwise, there were no problems encountered installing and booting the five (5) R630 management servers, 20 R730 storage servers, and the 320 M630 compute blades. Two (2) additional R630 management servers were removed from the J-AZ cluster and installed in the J-IU cluster before the J-AZ was packed up for shipment to Arizona where J-AZ will serve as the test and development resource for Jetstream.

The Jetstream-TACC (J-TACC) Production cluster was delivered to the TACC data center. No problems were encountered installing and booting the seven (7) R630 management servers, 20 R730 storage servers, and the 320 M630 compute blades.

Table 20. Detailed timeline of the delivery and acceptance tests for Jetstream.

Item	Jetstream - IU	Jetstream - TACC	Integrated Functions
Purchase Order	7/29/2015	7/29/2015	
System arrival	10/19/2015	10/16/2015	
System boot	11/11/2015	11/03/2015	
Basic system functionality pass (including XSEDE Integration)	1/14/2016	11/05/2015	
Performance pass	11/30/2015	02/24/2015	
Friendly User Mode			8/28/2015 on Jetstream-AZ test system
Early Operations	2/10/2016	3/3/2016	3/10/2016
Galaxy available for production use	4/15/2016	4/15/2016	
14-day system stability pass	4/20/2016	4/20/2016	
Galaxy correct function and speed test pass			4/20/2016
SEAGrid correctness test pass			4/15/2016
SEAGrid stability test pass			4/29/2016

11. Appendix II. Detailed hardware specifications and hardware performance test explanations

11.1. Hardware

This hardware description format is based on the format specified in http://www.nsf.gov/pubs/2006/nsf0605/nsf0605.jsp (Benchmarking Information Referenced in the NSF 11-511 "High Performance Computing System Acquisition: Towards a Petascale Computing Environment for Science and Engineering")

11.1.1. System topology

- The JI and JT production clusters consists of 7 PE R630 management servers, 20 (4) PE R730 storage servers and 320 PE M630 compute blades. The PE R630 management servers are configured with dual Intel 2.5 GHz, 120W, Xeon E5-2680v3 "Haswell" chips, 64 GB RAM, dual 400GB SSD system devices and are wired directly into the Dell Force10 (F10) S6000 spine network switch.
- The PE R730XD storage servers are configured with dual Intel E5-2680v3 "Haswell" chips, 64 GB DDR4 RAM, dual 200GB SSD system devices, and 12 – 4 TB Near-Line Serial Attached SCSI (NL-SAS) storage disks and are wired directly into the Dell Force10 S6000 spine network switch.
- The PE M630 blade servers are installed in a PE M1000 blade enclosure and are configured with dual Intel E5-2680v3 "Haswell" chips, 128 GB RAM, and dual 1 TB NL-SAS disk drives and are wired into the Dell PE MXL1000 chassis switches which then uplink into the Dell Force10 S6000 spine switch producing a two-to-one oversubscribed Fat-Tree topology.

11.1.2. Memory boards, sections, and/or banks

- Each M630, R630, and R730 has 24 DDR4 DIMM slots running at 2133MT/s

11.1.3. Memory size

- Each M630 compute blade has eight (8) 16 GB RDIMM running at 2133MT/s for a total of 128 GB.
- Each R630, R730 management, storage server respectively has eight (8) 8 GB RDIMM running at 2133MT/s for a total of 64 GB.

11.1.4. CPU manufacturer, model, and speed

- Each M630, R630, and R730 are populated with dual Intel Xeon E5-2680v3, 12-Core 2.5 GHz, 2133MHz bus with 30MB L3 cache, 12x256KB L2 cache.

11.1.5. *Speed of the memory and memory bus (if applicable)*

- Each of the M630, R630, and R730 utilize DDR4 memory running at 2133MT/s.

11.1.6. *I/O boards and bus interfaces*

- M630: internal RAID controller, Intel QPI @ 9.6 GT/s.
- R630: two (2) PCIe Gen3x16 slots, one (1) PCIe Gen3 x8 slot, dedicated RAID card; Intel QPI @ 9.6 GT/s.
- R730: two (2) PCIe Gen3x16 slots, four (4) PCIe Gen3 x8 slot, dedicated RAID card; Intel QPI @ 9.6 GT/s.

11.1.7. *HBAs, Network Interface Cards and TCO Offload Engine (TOE) cards including firmware*

- None.

11.1.8. *Network adapters, including firmware*

- M630: Intel X710 dual port, 10 Gbps, version 1.3.38, firmware-version: 4.25 0x8000143f 0.0.0.
- R630, R730: Intel X710 quad port, 10 Gbps, version 1.3.38, firmware-version: 4.25 0x8000143f 0.0.0.

11.1.9. *All communications hardware, including private channels*

- Dual Dell Networking MXL 10/40 Gbps Ethernet blade switches (leaf).
- Dell Force10 S6000 10/40 Gbps Ethernet top of rack switch and spine.
- Sixteen M630 blades connect via bonded 2 x 10 Gbps links to the two Dell MXL switches in each blade chassis (with virtual link trunking enabled) which each uplink to the Top-of-Rack (ToR) Dell Force10 (F10) S6000 switches via two bonded 40 Gbps links resulting in 2:1 over-subscription to the blades. Each ToR S6000 then connects via 2 x 40 Gbps links into each of the two Spine S6000 switches. The two spine F10 S6000 are cross linked at 3x40 Gbps for 120 Gbps aggregate. Each F10 S6000 spine switch is then uplinked to the data center's Science DMZ via 2 x 40 Gbps uplinks for a total of 4 x 40 Gbps. The R630 management and R730 storage nodes link into the F10 S6000 spine switch via dual bonded 10 Gbps links.
- Dell N3048 1 Gbps management Ethernet switch provide out-of-band management control of the overall system.

11.1.10. *RAID hardware including disks, cache, firmware, channels, GBICS and interfaces*

- M630: PERC H330 RAID controller; dual 1 TB 7.2K RPM NL-SAS 6 Gbps 2.5in Hot-plug system devices.

- R630: PERC H330 integrated RAID controller; dual 400 GB Solid State Drive (SSD) SATA Mix Use MLC 6 Gbps 2.5in Hot-plug system devices.
- R730: PERC H730P integrated RAID controller, 2 GB cache; dual 200 GB Solid State Drive SATA Mix Use MLC 6 Gbps 2.5in Flex Bay system devices; twelve 4 TB 7.2K RPM NL-SAS 6 Gbps 3.5in Hot-plug Hard storage devices.

11.1.11. Fibre Channel switches, if used

- None

11.1.12. Any other hardware used as part of the benchmark configuration

- Benchmarks were run from an NFS mounted file system exported from the respective cluster's management server.

11.2. Software

11.2.1. Operating system, including all tunable parameters and their values

- M630: CentOS 7.1.1503 w/ kernel 3.10.0-229.el7.x86_64, stock.
- VM: CentOS 7.2.1511 w/ kernel 3.10.0-229.el7.x86_64, stock.
- MXL: 9.9.9.0.0
- S6000-ON: 9.8.0.0p9

11.2.2. BIOS tunable parameters and their values

- Firmware
  - M630: 2.10.10.10, build 49, 04/06/2015 09:05:28.
  - R630: 2.02.01.01, build 92, 09/15/2014 09:45:31.
  - R730: 2.02.01.01, build 92, 09/15/2014 09:45:31.
- BIOS
  - M630: 1.1.10, default performance values.
  - R630: 1.0.4, default performance values.
  - R730: 1.2.10, default performance values.

11.2.3. Network drivers

- Intel i40e, version 1.3.47, firmware 4.41 0x8000186a 16.5.20.

11.2.4. Network stacks, including TOEs

- Standard Linux network stack.

11.2.5. I/O drivers

- N/A

11.2.6. *File system software and/or volume manager*

- xfs for local file systems.
- Ceph 0.94.5-0.el7 (Hammer) for block and object storage.

11.2.7. *Compiler and libraries, including I/O and MPI libraries*

- Intel compilers, version 15u3, Intel MPI version 5.0.3p-048

11.2.8. *All patches and bug fixes*

- CentOS 7.2.1511 with patches up to date as of the Performance Pass date identified in tables above.

11.2.9. *Any additional software used as part of the benchmark configuration*

- qemu-kvm-1.5.3-105.el7_2.1
- libvirt-daemon-kvm-1.2.17-13.el7_2.2
- OpenStack 2015.2.0 (Liberty)

12. Appendix III. Acceptance test criteria

The Project Execution Plan (PEP) between Indiana University and the National Science Foundation stipulate the acceptance criteria for Jetstream.

The purpose of the acceptance testing is to ensure that the system as implemented is the system described in the original proposal as modified by a scope of work change document. The following acceptance criteria demonstrate the functionality of Jetstream based exclusively on the terms of NSF Request for Proposals, the original proposal by IU and its partners, and the scope of work change document submitted to the NSF as a supplement to the original proposal. If completed successfully these tests will comprehensively demonstrate that the computational resource satisfies the capabilities of the Jetstream system that Indiana University and its subcontractors have been contracted to integrate and deliver.

IU and NSF retain the right by mutual agreement to change these tests should one or more prove not informative, or if the software underlying the tests proves itself to be faulty in terms of demonstrating the capabilities of Jetstream.

12.1. Basic hardware performance

Jetstream is a first-of-a-kind acquisition and implementation for the NSF and for the NSF-funded national cyberinfrastructure. It is more a system implementation than a hardware implementation (as contrasted, say, to earlier systems such as Ranger, Kraken, or FutureGrid). However, it makes sense to have some basic hardware performance tests as the first step in the acceptance testing of Jetstream. These criteria are, in a sense, prerequisites for other tests that verify the functionality of the system. These tests are primarily performance tests – the doing of a specific activity.

12.1.1. Single Node Performance

High-Performance Linpack (HPL): Single node Linpack performance will achieve 80% of the peak floating-point performance for a problem size that uses at least half of the on-node memory. (Measurements will be rounded to nearest %).

STREAM: Single node OpenMP threaded STREAM performance will be at least 65 GB/s (aggregate across the node). (Measurements will be rounded to nearest 1 GB/s).

10 Gigabit Ethernet Bandwidth: the 10 GigE interface on each node will achieve at least 1 GB/s for large-message point-to-point transfers (Measurements will be rounded to the nearest 0.1 GB/s).

12.1.2. File System and Storage Benchmarks

The system will achieve a minimum of 200 MB/s data transfer rate for data reads and a minimum of 100 MB/s writes from within a virtual machine from/to the block storage. (Measurements will be rounded to the nearest MB/s).

12.1.3. System Reliability Tests

System reliability will be tested by operating the system during the friendly user mode with uptime of at least 95% for a period of 14 days. Appendix 1 of the PEP describes the rationale for a 14-day reliability test.

Neither the solicitation nor our proposal included any terms regarding Mean Time Between Failures (MTBF), so MTBF is not included as part of the acceptance criteria. However, we can place a lower bound on MTBF from the system reliability metrics. 95% uptime implies that the system won't be down more than 36 hours per month.

12.2. Provide "self-serve" academic cloud services

The full text of the capability described in Section 2 of the PEP is 'Provide "self-serve" academic cloud services, enabling researchers or students to select a VM image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing. Authentication to this "self-serve" environment will be via Globus.' Implicit in the sense of the words 'cloud services' is that the two production components of Jetstream function as parts of an integrated whole. There are both capability and capacity issues to providing a cloud environment.

Much of the description of Jetstream as a cloud resource describes capabilities so the tests of these aspects are 'capability' tests, and a first of a kind system the test will consist simply of demonstrating the following functions:

- An authorized and knowledgeable user will be able to authenticate to the Jetstream user interface (which uses Globus as the mechanism for verification of credentials).
    - After so doing, an authorized and knowledgeable user will be able to launch a virtual machine from a menu of pre-packaged VMs on the production hardware located in Indiana or Texas.
    - After so doing, an authorized and knowledgeable user will be able to quiesce a VM image running on production hardware in Indiana or Texas, move it from one production system to another, and reactivate said VM.
- An authorized and knowledgeable user can create and access persistent cloud storage on the Indiana or Texas production hardware
- An authorized and knowledgeable user can modify a preexisting VM image and manually store that VM image to one of the production locations within Jetstream.

There is a capacity (load) goal that can be derived from the proposal as well. Working backwards from the final budget and configuration and the statements in the original proposal limiting the amount of oversubscription that would be permitted on Jetstream, and VM configurations, we can create a metric of the minimum number of active VMs that Jetstream should support: 640. (This is based on 640 nodes in the system, with the largest VMs to be supported on Jetstream taking a full node) This leads to the following system capacity test:

- Jetstream will support a minimum of 640 VMs operating simultaneously.

12.3. Host persistent science gateways

The full text of the capability described in Section 2 of the PEP is 'Host persistent Science Gateways. Jetstream will support persistent science gateways, including the capability of hosting persistent science gateways within a VM when the nature of the gateway is consistent with operation within a VM. Galaxy will be one of the initial science gateways supported.'

This is a 'capability' and functionality test, and a first of a kind system the test will consist of:

- The Galaxy bioinformatics gateway is installed and will operate a demonstration workflow providing correct results, based on comparison with output results from a known reference installation. The job will complete within 25% of the time required to complete an analysis running on an equivalent system.
- One other exemplar science gateway that is known to function properly in other XSEDE-supported gateway hosting environments will function and remain reliable to within 2% of the overall system availability achieved during system reliability tests during a 14-day test period. (E.g. if the system turns out to be available with an uptime of 96%, the gateway used to test this criterion will be available 96% - 2% or 94%). The test period may be contemporaneous with the overall system test period or done at some other time. The critical metric here is that Gateway Availability track overall availability within a delta of 2% of total potential system uptime.

12.4. Data movement, storage and dissemination

The full text of the capability described in Section 2 of the PEP is 'Data movement, storage and dissemination.'

- *'Jetstream will support data transfer with Globus Connect.*
- *Users will be able to store VMs in the Indiana University persistent digital repository, IUScholarWorks (scholarworks.iu.edu) and obtain a Digital Object Identifier (DOI) that is associated with the VM stored.'*
- *The performance characteristics of the storage system are verified through item 12.1.2. Globus Connect is a service offered by a partner organization that contains a set of performance characteristics that are well understood, and not affected by this solicitation. The first item above becomes a functionality test:*
- *An authorized and knowledgeable user can select a file to which they have rights on a system outside Jetstream, and move that file and save it on storage on Jetstream (with the condition that the file size is within the storage quota set for their use on Jetstream).*
- *An authorized and knowledgeable user can select a file to which they have rights on Jetstream, and move that file and save it on storage to a system on which that user has rights and which is accessible from open public networks (with the condition that the file size is within the storage quote set for their use on Jetstream).*

The second feature described above is again a capability test, satisfied by the following:

- An authorized and knowledgeable user can successfully save a VM previously stored to disk storage on Jetstream into a format supported by DSpace, upload that file to IU Scholarworks.iu.edu, and using the existing online forms submit that document for publication via IUScholarWorks. Subsequent to that, provided the relevant and required information has been provided by the user, the VM will appear in IUScholarWorks and the user will receive a DOI identifier for that object. Note: This is a "human in the loop" process and may take days from upload and submission to publication and receipt of DOI. Email transactions may be required beyond the initial submission.

12.5. Provide virtual Linux desktop services delivered from Jetstream to tablet devices

The full text of the capability described in Section 2 of the PEP is

'Provide virtual Linux desktop services delivered from Jetstream to tablet devices. This service is aimed to increase access to Jetstream for users at institutions with limited resources including small schools, schools in EPSCoR states, and Minority Serving Institutions.'

This test is a functionality test, with some time constraints. This feature will be satisfied by the following:

- An authorized and knowledgeable user can access Jetstream from a tablet device, and load a virtual Linux desktop configured in a way that allows the user to access Jetstream services.

13. Appendix IV. Hardware acceptance test methodology and results

13.1. Basic hardware performance

13.1.1. Single node performance tests

- Single node performance benchmarks were run on all M630 compute servers on both the Jetstream-IU and Jetstream-TACC clusters. The Jetstream-AZ system was previously accepted.

13.1.2. HPL

- The theoretical peak performance for the Dell M630 server is 806.4 GFLOPS. For a node to pass acceptance, it must achieve 80% of this value or 645.1GFLOPS on the HPL. Measurements will be rounded to the nearest 1%.
- HPL was run as part of the HPCC benchmark suite. No modifications to the source code were made. It was compiled with the Intel compiler version 15.0.3 with options "-O3 -DRA_SANDIA_OPT2 -mP2OPT_hlo_loop_intrinsic=F."
- The performance target was achieved. The average performance across all tested servers was 697 GFLOPS (86% of theoretical peak performance) for JI and 701 (87%) GFLOPS for JT.

13.1.3. STREAM

- The memory performance target for an M630 node is 65 GB/s rounded to the nearest 1 GB/s.
- STREAM was run as a separate benchmark with no modifications to the source code. It was compiled with the Intel compiler version 15.0.3 with options "-O3 –xCORE-AVX2 –openmp."
- The performance target was achieved for STREAM. The average STREAM Triad performance across all tested servers was 100.5 GB/s for JI and 113.1 for JT.

13.1.3.1. Ethernet bandwidth

- Each node will need to demonstrate 1 GB/s rounded to the nearest 0.1 GB/s across its 10 Gbps interfaces.
- Iperf, with default settings, was used to measure Ethernet bandwidth and run between a management node and all M630 compute, R730 storage, and R630 management servers.
- For the Ethernet Bandwidth benchmark, the performances target was achieved. The average performance across all tested servers was 1.1 GB/s for JI and 1.2 GB/s for JT.

13.1.4. *File system and storage performance tests*

- A minimum read performance of 200 MB/s and a minimum write performance of 100 MB/s from within a virtual machine to block storage. Measurements will be rounded to the nearest MB/s.
- Read/write I/O performance was measured via the dd Linux utility to/from an OpenStack Cinder block device mounted within a running VM instance. A file size of 2 GB was used with a block size of 1 MB.
- The performance targets were achieved for file system and storage. The write performance for a single VM was 359 MB/s for JI and 108 MB/s for JT. The read performance for a single VM was 244 MB/s for JI and 210 MB/s for JT.

13.1.5. *System stability and uptime performance tests*

The system must maintain a continuous 95% availability for a period of 14 days.

For the Jetstream system, i.e. JI and JT running as an integrated entity, the system must be up, running stably, and available for users to engage in their routine research activities. The Jetstream system was up, running stably, and 100% available for daily usage for over 14 days starting April 6, 2016 through April 21, 2016.

MTBF requirements are not applicable to the test environment but the system operated continuously for over 28 days.

13.2. Integrated cloud operations

The inherent value of Jetstream is not in its hardware components; but rather, in the integration of the various software and hardware parts into its whole. Metrics designed to demonstrate the achievement of this are listed below.

13.2.1. *Provide "self-service" academic cloud services*

On a routine and daily basis, users of the Integrated Jetstream system are:

- Authenticating via GlobusAuth to the Jetstream Atmosphere user interface. As of April 21, 2016, 159 users from 68 distinct projects have access to Jetstream.
- Launching virtual machine instances from the menu of pre-packaged VM images installed in Jetstream's libraries.
- Suspending a running instance to as disk image.
- Restart a quiesced image on the production portion of Jetstream different from the one it was initially started on.
- Creating and accessing persistent cloud storage (volumes).

Jetstream has also demonstrated the ability to instantiate and sustain more than 640 VMs operating simultaneously. The Jetstream team has performed multiple quality assurance tests

during the early operations phase to demonstrate the system can exceed the minimum value. Jetstream as an integrated system and each site individually have exceeded the target value: 998 VMs were running simultaneously at IU on April 15, 2016 and 832 VMs were running simultaneously at TACC on March 7, 2016. On April 29, 2016 a combined 1217 instances were running simultaneously between the two sites.

13.2.2. Data movement, storage, and dissemination

- Authorized users have transferred files into and out of Jetstream utilizing Globus Connect Personal from within a running instance. In addition, users can leverage their own desired transfer tools such as SFTP, iRODS, and HTTP protocols.
- Users have been able to suspend a running instance, save it to disk, upload it to IU scholarworks.iu.edu, and using the existing online forms, submit that document for publication via IUScholarWorks. E.g. CentOS 7 (7.2) Development; Fischer, Jeremy; Stewart, Craig; DOI: doi:10.5967/P93W2M; URI: http://hdl.handle.net/2022/20772

13.2.3. Provide virtual Linux desktop services delivered from Jetstream to tablet Devices

- Authorized users have accessed Jetstream from a tablet device and loaded a virtual Linux desktop configured in a manner that allowed them to access Jetstream services. This functionality was tested using the RealVNC viewer app on an iOS device.

14. Appendix V. Detailed results of Galaxy validation tests and performance analysis.

Jetstream has shown the ability to support persistent science gateways
The Galaxy bioinformatics gateway has been installed and became operational on April 15, 2016. A known Galaxy workflow was executed and provided the correct results within the specified performance parameters.

A known workflow BWA MEM https://usegalaxy.org/u/jxtx/h/e-coli-pacbio-bwa was run on the Indiana University Mason cluster, TACC's Stampede system, and on Jetstream.

BWA-MEM is a new alignment algorithm for aligning sequence reads or long query sequences against a large reference genome. This workflow maps 20x coverage PacBio reads to the E Coli K12 reference sequence. The acceptance test as described in the PEP and quoted in section 12.3 indicates this workflow on Jetstream should be no more than 25% slower than the reference system. To this end, we directly compared Jetstream and Mason (at IU): the total execution time on Jetstream was 196 seconds as compared to 471 seconds on Mason. The workflow on Jetstream took a fraction of the time as compared to Mason, 41%. If one normalizes based on clock speed (2.5 GHz as compared to 1.86 GHz) the Jetstream execution is 56%.

The output of the three tests are included below for reference:

Jetstream Output:

```
[bwa_index] Pack FASTA... 0.06 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.98 seconds elapse.
[bwa_index] Update BWT... 0.04 sec
[bwa_index] Pack forward-only FASTA... 0.03 sec
[bwa_index] Construct SA from BWT and Occ... 0.46 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa index localref.fa
[main] Real time: 1.874 sec; CPU: 1.582 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa mem -t 10 -v 1 -x pacbio localref.fa /jetstream/iu-
scratch0/main/jobs//12479107/inputs/dataset_15262597.dat
[main] Real time: 196.026 sec; CPU: 1895.285 sec
```

Mason Output:

```
[bwa_index] Pack FASTA... 0.06 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 1.80 seconds elapse.
[bwa_index] Update BWT... 0.05 sec
[bwa_index] Pack forward-only FASTA... 0.04 sec
[bwa_index] Construct SA from BWT and Occ... 0.90 sec
[main] Version: 0.7.12-r1039
```

[main] CMD: bwa-0.7.12/bwa index localref.fa
[main] Real time: 131.394 sec; CPU: 2.878 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa-0.7.12/bwa mem -t 10 -v 1 -x pacbio localref.fa pacbio.fastq
[main] Real time: 471.434 sec; CPU: 4384.194 sec

Stampede Output (6 threads vs 10 threads for prior runs):

[bwa_index] Pack FASTA... 0.03 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 1.18 seconds elapse.
[bwa_index] Update BWT... 0.03 sec
[bwa_index] Pack forward-only FASTA... 0.04 sec
[bwa_index] Construct SA from BWT and Occ... 0.50 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa index localref.fa
[main] Real time: 2.476 sec; CPU: 1.784 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa mem -t 6 -v 1 -x pacbio localref.fa /galaxy-repl/main/files/015/262/dataset_15262597.dat
[main] Real time: 410.749 sec; CPU: 2375.358 sec

15. Appendix VI. Example letter of commitment to a Principal Investigator who has requested commitment from Jetstream in support of another NSF proposal

INDIANA UNIVERSITY
OFFICE OF THE VICE PRESIDENT FOR
INFORMATION TECHNOLOGY AND
CHIEF INFORMATION OFFICER

February 20, 2015

[redacted]

Dear [redacted]

On behalf of the Jetstream cloud computing initiative, I am writing this letter of commitment to collaborate with you as you develop your [redacted] for building and maintaining large dynamic databases for social science research as part of your proposal to NSF solicitation 15-532 (RIDIR).

Specifically, we will collaborate with your team to connect your [redacted]ystem to our Jetstream system in order to provide a broader array of computing resources to researchers using your tools. Jetstream, a significant initiative recently funded by NSF, will use virtual machine and cloud-computing resources to give researchers greater access to high performance computing that are generally not available to most social science researchers. 90% of the time on Jetstream is allocated through a peer-review process managed by the eXtreme Science and Engineering Discovery Environment. I am confident that you and users of your services will have no difficulty obtaining allocations of use of Jetstream through this process. 10% of the capacity of Jetstream is allocated at my discretion as Principal Investigator. I will use some of this discretionary allocation if needed to ensure that you, your colleagues, and your users are successful in carrying out this important work.

We are pleased to work with you on this important project.

Sincerely,

Craig A. Stewart, Ph.D.
Executive Director, Pervasive Technology Institute
Associate Dean, Research Technologies

2709 E. 10th Street, Suite 301 Bloomington, IN 47408 (812) 855-4717 http://ovpit.iu.edu
535 W. Michigan Street, IT 500 Indianapolis, IN 46202 (317) 278-3960

16. Appendix VII. Detailed results of SEAGrid validation tests

In order to verify correct functioning of SEAGrid on Jetstream we ran two runs of the same analysis doing ligand docking: one launched from Jetstream and executed on Comet; one launched from Jetstream and executed on Jetstream. Figure 24 below shows the former about to be launched. Below that are the std.out files from the job that ran on Comet and the job that ran on Jetstream.

This test verifies two things. First, as a gateway, SEAGrid on Jetstream successfully dispatches work to another XSEDE resource (Comet), as a gateway should. This test also verifies that Jetstream produces the same computational results as Comet on a sample computational task.

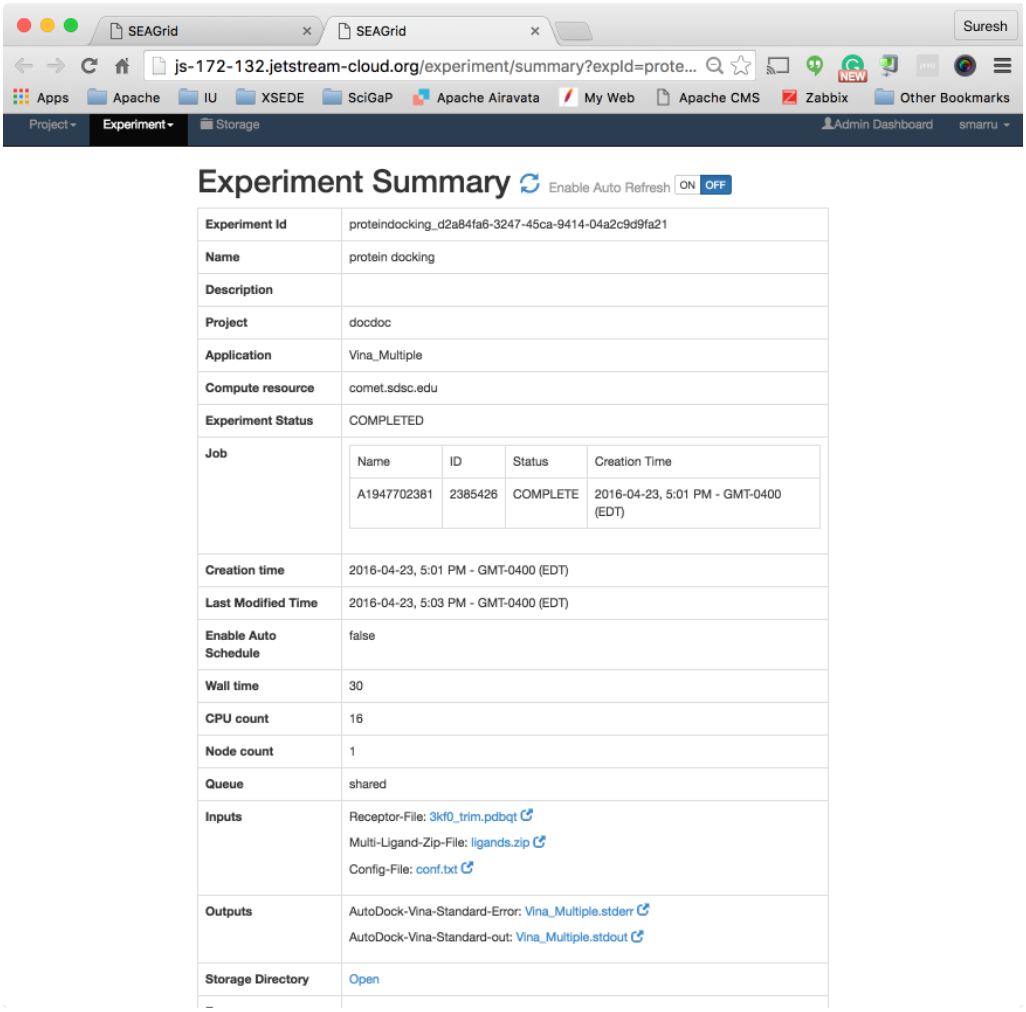


Figure 24. SEAGrid experiment summary showing gateway accepting a job on Jetstream to execute on Comet

16.1. Jetstream std.out from SEAGrid

```
Archive:  ligands.zip
  inflating: 1f1.pdbqt
  inflating: 4DX.pdbqt
Processing ligand 1f1
###############################################################
# If you used AutoDock Vina in your work, please cite:          #
#                                                               #
# O. Trott, A. J. Olson,                                        #
# AutoDock Vina: improving the speed and accuracy of docking    #
# with a new scoring function, efficient optimization and       #
# multithreading, Journal of Computational Chemistry 31 (2010)  #
# 455-461                                                       #
#                                                               #
# DOI 10.1002/jcc.21334                                         #
#                                                               #
# Please see http://vina.scripps.edu for more information.      #
###############################################################

Detected 24 CPUs
WARNING: at low exhaustiveness, it may be impossible to utilize all CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -1645829847
Performing search ...
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************
done.
Refining results ... done.

mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1         -4.9      0.000      0.000
   2         -4.4      2.933      4.414
   3         -4.3      1.354      2.039
Writing output ... done.
Processing ligand 4DX
###############################################################
# If you used AutoDock Vina in your work, please cite:          #
#                                                               #
# O. Trott, A. J. Olson,                                        #
# AutoDock Vina: improving the speed and accuracy of docking    #
# with a new scoring function, efficient optimization and       #
# multithreading, Journal of Computational Chemistry 31 (2010)  #
# 455-461                                                       #
#                                                               #
# DOI 10.1002/jcc.21334                                         #
#                                                               #
```

```
# Please see http://vina.scripps.edu for more information.      #
#################################################################

Detected 24 CPUs
WARNING: at low exhaustiveness, it may be impossible to utilize all CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -752121435
Performing search ...
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************
done.
Refining results ... done.

mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1         -3.2      0.000      0.000
   2         -3.2      1.491      2.684
   3         -3.1      1.963      3.459
Writing output ... done.
```

16.2. Comet std.out

```
Archive:  ligands.zip
  inflating: 1f1.pdbqt
  inflating: 4DX.pdbqt
Processing ligand 1f1
#################################################################
# If you used AutoDock Vina in your work, please cite:          #
#                                                               #
# O. Trott, A. J. Olson,                                        #
# AutoDock Vina: improving the speed and accuracy of docking    #
# with a new scoring function, efficient optimization and       #
# multithreading, Journal of Computational Chemistry 31 (2010)  #
# 455-461                                                       #
#                                                               #
# DOI 10.1002/jcc.21334                                         #
#                                                               #
# Please see http://vina.scripps.edu for more information.      #
#################################################################

Detected 24 CPUs
WARNING: at low exhaustiveness, it may be impossible to utilize all CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -1645829847
Performing search ...
```

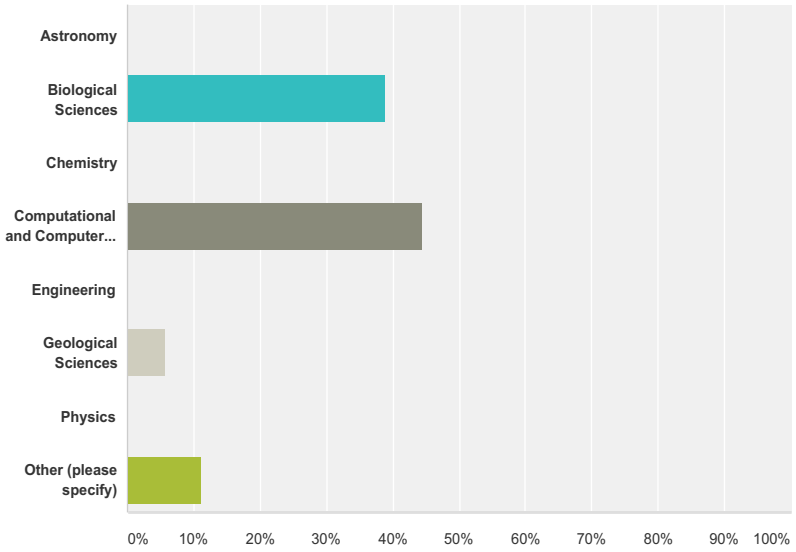
```
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************
done.
Refining results ... done.

mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1         -4.9      0.000      0.000
   2         -4.4      2.933      4.414
   3         -4.3      1.354      2.039
Writing output ... done.
Processing ligand 4DX
###############################################################
# If you used AutoDock Vina in your work, please cite:          #
#                                                               #
# O. Trott, A. J. Olson,                                        #
# AutoDock Vina: improving the speed and accuracy of docking    #
# with a new scoring function, efficient optimization and       #
# multithreading, Journal of Computational Chemistry 31 (2010)  #
# 455-461                                                       #
#                                                               #
# DOI 10.1002/jcc.21334                                         #
#                                                               #
# Please see http://vina.scripps.edu for more information.      #
###############################################################

Detected 24 CPUs
WARNING: at low exhaustiveness, it may be impossible to utilize all CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -752121435
Performing search ...
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************
done.
Refining results ... done.

mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1         -3.2      0.000      0.000
   2         -3.2      1.491      2.684
   3         -3.1      1.963      3.459
Writing output ..
```

17. Appendix VIII. Initial Jetstream feedback

Jetstream Early Ops Assessment

Q1 What is your primary field of research?

Answered: 18 Skipped: 1

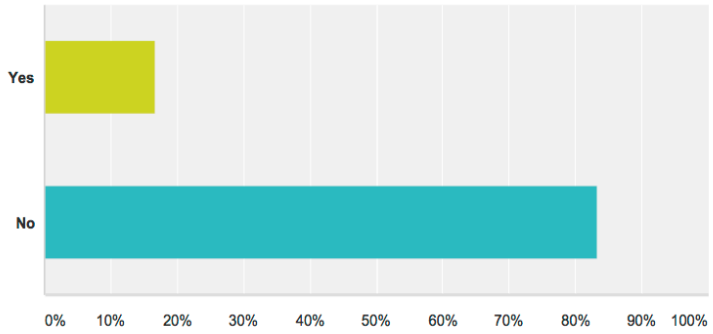


Answer Choices	Responses	
Astronomy	0.00%	0
Biological Sciences	38.89%	7
Chemistry	0.00%	0
Computational and Computer Sciences	44.44%	8
Engineering	0.00%	0
Geological Sciences	5.56%	1
Physics	0.00%	0
Other (please specify)	11.11%	2
Total		18

#	Other (please specify)	Date
1	Research Support	4/27/2016 8:13 PM
2	canine genomics	4/27/2016 4:56 PM

Q2 Are you an XSEDE campus champion?

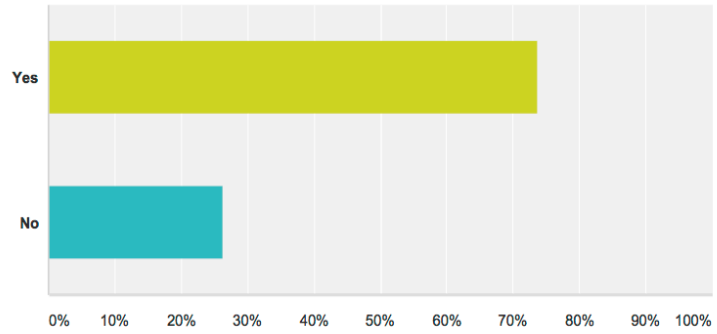
Answered: 18 Skipped: 1



Answer Choices	Responses	
Yes	**16.67%**	3
No	**83.33%**	15
Total		**18**

Q3 Have you tried to use Jetstream yet?

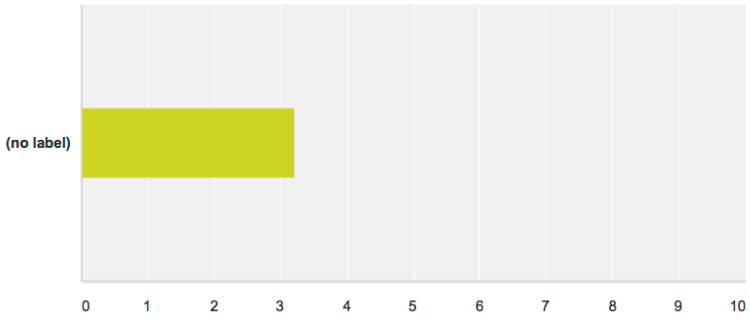
Answered: 19 Skipped: 0



Answer Choices	Responses	
Yes	73.68%	14
No	26.32%	5
Total		19

Q5 If you have used Jetstream, please rate your experience on a scale from 1-5, with "1" being "extremely dissatisfied," and "5" being "extremely satisfied."

Answered: 14 Skipped: 5



	1-Extremely Dissatisfied	2	3- Neither Satisfied or Dissatisfied	4	5- Extremely Satisfied	Total	Weighted Average
(no label)	7.14% 1	14.29% 2	28.57% 4	50.00% 7	0.00% 0	14	3.21

18. Appendix IX. Outreach activities

Table 21. Outreach activities

Username	Date	Event Title	Conference Name/Location	Description
bahalloc	12/4/2014	Presentation on Jetstream and Wrangler	NSF Workshop on High Performance Distributed Computing and Polar Sciences / Rutgers University New Brunswick NJ	Justin Miller from Indiana University's (IU) HPC group presented a new NSF cloud system Jetstream and NSF data science infrastructure called Wrangler
dyhancoc	1/28/2015	Cyberinfrastructure Begins at Home	SPXXL Winter Workshop (IBM/Lenovo HPC User Group)	A look at the history of IU cyberinfrastructure and new projects such as XSEDE and Jetstream
dyhancoc	4/15/2015	Jetstream: A science & engineering cloud	56th HPC User Forum / Norfolk VA	Jetstream project outreach to inform the HPC community at large
jeremy	4/9/2014	Introduction to XSEDE	Wittenberg CI Days / Springfield, OH	Talk and panel on XSEDE and other national CI
jeremy	5/11/2015	XSEDE / ACI-REF Meeting	Clemson, SC	Met with Clemson and ACI-REF staff to discuss merging efforts
jeremy	6/16/2015	Jetstream: A Distributed Cloud Infrastructure for Underresourced higher education communities	The Science of Cyberinfrastructure: Research Experience Applications and Models (SCREAM-15) / Portland OR	Jeremy Fischer gave the presentation.
jeremy	10/20/2015	IU Statewide IT Conference - NSF Jetstream Science and Engineering Cloud	Statewide IT Conference / IUB - Indiana Memorial Union	Jetstream - a national science and engineering cloud
jeremy	11/2/2015	Southern Partnership in Advanced Networking (SPAN) - Workshop 2 - Jetstream	Southern Partnership in Advanced Networking - Workshop 2 / Huntsville AL	Workshop for southern universities supporting research and teaching activities
jeremy	1/6/2016	Overview of Jetstream for Earth Sciences	2016 ESIP Winter Meeting / Washington DC	Presentation to researchers and IT staff
jeremy	3/9/2016	Southern Partnership in Advanced Networking (SPAN) - Workshop 3 - Jetstream	University of Central Florida Orlando FL	Presented overview of Jetstream for southeastern US IT and research staff
jeremy	3/25/2016	XSEDE Gateway Communities Call	Teleconference	
robping	5/18/2015	Jetstream Overview - EGI 2015	European Grid Infrastructure / Lisbon Portugal	Vas Vasiliadis presented a talk on Jetstream.
seiffert	4/7/2015	Jetstream Lightning Talk	GlobusWorld 2015 / Argonne National	Presentation on Jetstream capabilities and purpose

Username	Date	Event Title	Conference Name/Location	Description
			Laboratory	
stewart	11/20/2014	Jetstream: A national science and engineering cloud	SC14 - New Orleans, LA	Present the Jetstream system as a national CI resource under XSEDE
stewart	12/1/2014	Jetstream: A science & engineering cloud.	Polar Workshop / Bloomington, IN	Present the Jetstream system as a national CI resource under XSEDE
stewart	2/1/2015	Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online	IU School of Public and Environmental Affairs	Presentation to the IU School of Public and Environmental Affairs
stewart	2/10/2015	Jetstream: A national science & engineering cloud	IUPUI Campus / Indianapolis, IN	Presentation to the IUPUI Faculty Coucil Information Technology Subcommittee
stewart	2/24/2015	Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online	Indiana University /Bloomington, IN	Presentation to Microsoft, Inc. Visiting Group
stewart	2/27/2015	Jetstream: A national science & engineering cloud	Indiana University /Bloomington, IN	Presentation to the IU Bloomington Faculty Council Information Technology Subcomittee
stewart	5/22/2015	Jetstream EPSCoR Outreach activity	2015 KY EPSCoR Annual Conference / Lexington KY	Craig Stewart was invited plenary lunch speaker
turnerg	7/27/2015	Clouds are Forming	XSEDE15 / St. Louis MO	Panel discussion
stewart	7/28/2015	Stewart C.A. 2015. Jetstream - A self-provisioned scalable science and engineering cloud environment. Presentation. XSEDEâ€™15	XSEDE15 / St. Louis, MO	Stewart C.A. 2015. Jetstream - A self-provisioned scalable science and engineering cloud environment. Presentation. XSEDEâ€™15 July 26 - 30 2015. St. Louis MO USA. http://hdl.handle.net/2022/20338
stewart	7/29/2015	Jetstream Overview XSEDE15 Panel - New and emerging US cyberinsfrastructure resources	XSEDE15 / St. Louis MO	Presentation as part of a plenary panel about Jetstream
stewart	10/20/2015	Cyberinfrastructure for Research: New Trends and Tools.	University of Vermont, Burlington VT	Stewart C.A. 2015. Presentation: http://hdl.handle.net/2022/20414
stewart	10/21/2015	Cyberinfrastructure for Research: New Trends and Tools.	University of Michigan	Stewart Craig A. Presentation: http://hdl.handle.net/2022/20445
stewart	10/22/2015	Cyberinfrastructure for Research: From campus growth to national trends.	Cyberinfrastructure Days / Michigan State University	Stewart Craig A. 2015. Presentation: http://hdl.han
stewart	11/17/2015	Demonstration of Jet Stream	SC15 Austin, TX	Demonstration of jet stream at SC 15

Username	Date	Event Title	Conference Name/Location	Description
turnerg	11/17/2015	Virtualization and Clouds in HPC: Motivation, Challenges & Lessons Learned	Austin, TX	Panel discussion
stewart	1/27/2016	Keynote talk: Exascale on what dimension and why?	SPPEXA Annual Program Meeting / Munich Germany	Keynote presentation by Craig A. Stewart
turnerg	3/8/2016	IU School of Informatics and Computing Educational Activity - Jetstream	IUB - School of Informatics and Computing	HPS staff described Jetstream's hardware and software architecture use cases and operational aspects of a functioning cloud system
turnerg	3/1/2016	IU Policy and Security Office Educational Activity-Jetstream	IUPUI - Informatics & Communications Technology Complex	HPS staff presented an architectural description of Jetstream to Indiana University's Security and Policy offices' monthly ChalkTalk
dyhancoc	4/12/2016	Jetstream Overview - IU/ZIH Collaboration	Virtual workshop with Technical University in Dresden (ZIH)	Jetstream overview and relevance to ZIH OpenStack project
mvaughn	4/16/2016	Jetstream: Adding Cloud-base computing to the National Cyberinfrastructure	HPC User Forum	Matt Vaughn presents to the HPC User Forum on Jetstream
jomlowe	4/27/2016	Deploying OpenStack for The National Science Foundation's Newest Supercomputers	OpenStack Summit 2016, Austin, TX	Mike Lowe and Bob Budden describe Jetstream and Bridges - architecture, installation, and use cases

19. References

[1] National Science Foundation, "NSF Solicitation 14-536." [Online]. Available: http://www.nsf.gov/pubs/2014/nsf14536/nsf14536.htm. [Accessed: 19-Apr-2016].

[2] NSF, "Natioal Science Foundation FY 2015 Performance and Financial Highlights." [Online]. Available: http://www.nsf.gov/pubs/2016/nsf16003/nsf16003.pdf. [Accessed: 19-Apr-2016].

[3] D. Hart, "Personal Correspondance." Boulder, CO, 2016.

[4] M. Dahan, "Personal Correspondance." Austin, Texas, 2016.

[5] D. Lifka, I. Foster, S. Mehringer, M. Parashar, P. Redfern, C. A. Stewart, and S. Tuecke, "XSEDE Cloud Survey Report," Indiana University, Bloomington, IN, 2013.

[6] P. B. Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," *Libr. Trends*, vol. 57, no. 2, pp. 280–299, 2008.

[7] M. McRobbie, G. Fox, D. Gannon, M. J. Palakal, and C. A. Stewart, "Extensible Terascale Facility (ETF): Indiana-Purdue Grid (IP-Grid)," Indiana University, Bloomington, IN, 2006.

[8] C. A. Stewart, D. S. Katz, D. L. Hart, D. Lantrip, D. S. McCaulay, and R. L. Moore, "Technical Report: Survey of cyberinfrastructure needs and interests of NSF-funded principal investigators," 2011.

[9] C. A. Stewart, R. Knepper, A. Grimshaw, I. Foster, F. Bachmann, D. Lifka, M. Riedel, and S. Tuecke, "Campus Bridging Use Case Quality Attribute Scenarios," 2012.

[10] C. A. Stewart, R. Knepper, A. Grimshaw, I. Foster, F. Bachmann, D. Lifka, M. Riedel, and S. Tueke, "XSEDE Campus Bridging Use Cases," p. 22, 2012.

[11] Committee on Network Science for Future Army Applications National Research Council, *Network Science*. Washington, DC: National Academies Press, 2005.

[12] "Network Workbench." [Online]. Available: http://nwb.cns.iu.edu. [Accessed: 24-Apr-2014].

[13] S. Wang, L. Anselin, B. L. Bhaduri, L. Nyerges, Timothy, N. R. Wilkins-Diehr, M. F. Goodchild, and A. Padmanabhan, "Year 3 Project Report – NSF SI2-SSI: CyberGIS Software Integration for Sustained Geospatial Innovation," 2013.

[14] K. Rodriguez, "UTSA launches first Open Compute Certification and Solution Laboratory in North America." [Online]. Available: http://www.eurekalert.org/pub_releases/2014-01/uota-ulf012814.php. [Accessed: 08-May-2014].

[15] Indiana University, "IUScholarWorks Repository." [Online]. Available: http://scholarworks.iu.edu/dspace. [Accessed: 07-May-2014].

[16] Globus, "Data Publication with Globus." [Online]. Available: https://www.globus.org/data-publication. [Accessed: 12-May-2014].

[17] Pegasus, "Pegasus." [Online]. Available: http://pegasus.isi.edu/. [Accessed: 16-Apr-2014].

[18] Taverna, "Taverna." [Online]. Available: http://www.taverna.org.uk.

[19] UNICORE, "UNICORE." [Online]. Available: http://www.unicore.eu. [Accessed: 07-

May-2014].

[20] The Kepler Project, “The Kepler Project.” [Online]. Available: https://kepler-project.org/. [Accessed: 17-Jan-2012].

[21] “The Apache Software Foundation.” [Online]. Available: http://www.apache.org/. [Accessed: 07-May-2014].

[22] B. D. Voss Jung-Gribble, Diane., Stewart, Craig A., “The Leveraged Support Model,” 1996.

[23] B. Voss, C. Stewart, and S. Workman, “Measuring Quality, Cost and Value of IT Services (Long Version),” *Annual Quality Congress, 55th*, vol. Proceeding. American Society for Quality, Charlotte, NC, 2001.

[24] C. A. Stewart, T. M. Cockerill, I. Foster, D. Hancock, N. Merchant, E. Skidmore, D. Stanzione, J. Taylor, S. Tuecke, G. Turner, M. Vaughn, and N. Gaffney, “Jetstream - A self-provisioned, scalable science and engineering cloud environment,” in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, 2015.

[25] J. Fischer, S. Tuecke, I. Foster, and C. A. Stewart, “Jetstream: A Distributed Cloud Infrastructure for Under-resourced Higher Education Communities,” *SCREAM ’15 Proc. 1st Work. Sci. Cyberinfrastructure Res. Exp. Appl. Model.*, 2015.

[26] J. Towhs, K. Gaither, R. Roskies, N. Wilkins-Diehr, G. Peterson, C. Hempel, S. Lathrop, K. Gendler, J. Navarro, E. Brooks, and D. Stanzione, “XSEDE Quarterly Report for Program Year 5 Q3: January 1, 2016 - March 31, 2016,” 2016.

[27] D. Y. Hancock, M. R. Link, C. A. Stewart, and G. W. Turner, “Acceptance Test for Jetstream Test Cluster — Jetstream-Arizona (JA) Dell PowerEdge Test and Development Cluster,” Aug. 2015.

[28] “Poweredge Server.” [Online]. Available: http://www.dell.com/us/business/p/servers.aspx?ST=dell poweredge server&dgc=ST&cid=294374&lid=5630222&acd=1230921533720565&ven1=sDsaJ1iR0&ven2=e&ven3=675303084151501206. [Accessed: 19-Apr-2016].

[29] “OpenStack.” [Online]. Available: https://www.openstack.org/. [Accessed: 19-Apr-2016].

[30] M. A. Davis, M. R. Douglas, M. L. Collyer, and M. E. Douglas, “Deconstructing a Species-Complex: Geometric Morphometric and Molecular Analyses Define Species in the Western Rattlesnake (Crotalus viridis),” *PLoS One*, vol. 11, no. 1, p. e0146166, 2016.

[31] “Cornel leads new National Science Foundation federated cloud project.” [Online]. Available: https://www.cac.cornell.edu/about/news/aristotle.aspx. [Accessed: 19-Apr-2016].

[32] “The Bro Network Security Monitor.” [Online]. Available: https://www.bro.org/. [Accessed: 19-Apr-2016].

[33] W. Application, A. Analysis, U. Bro, and S. I. I. R. Room, “Web Application Attack Analysis Using Bro Ids,” 2012.

[34] C. A. Stewart, R. Roskies, R. Knepper, R. L. Moore, J. Whitt, and T. M. Cockerill, “XSEDE Value Added, Cost Avoidance, and Return on Investment,” in *Proceedings of*

the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, 2015, pp. 23:1–23:8.